A picture containing text, tree, outdoor, resort

Description automatically generated

# **Keystroke Behavioral Biometrics Continuous Authentication Approach Using ML Algorithms**

**Generators** | **cys499** | **10/6/2023**

A picture containing shape

Description automatically generated

Information Technology Department

Faculty of Information Technology and Computer Science

Yarmouk University

Irbid, Jordan

*A picture containing shape

Description automatically generated*

*BSc Project :*

*IT Department :*

*Project ID:*

# **Keystroke Behavioral Biometrics Continuous Authentication Approach Using ML Algorithms**

**By**

**Mohammad AL Nawayseh (2019804016)**

**Nehad Dhoun , (2019904186)**

**Salma Jedayah , (2019904149)**

**Salsabeel Khaleel , (2019904179)**

Department of Information Technology

Faculty of Information Technology and Computer Sciences

Yarmouk University, Jordan

# **Intellectual Property Right Declaration**

# *This is to declare that the work under the supervision of* *Mohammad Rasool Al-Jaafrah having title “Keystroke Behavioral Biometrics Continuous Authentication Approach Using ML Algorithms*

# *” carried out in partial fulfillment of the requirements of Bachelor of Cybersecurity, is the sole property of the Yarmouk University and the respective supervisor and is protected under the intellectual property right laws and conventions. It can only be considered/ used for purposes like extension for further enhancement, product development, adoption for commercial/organizational usage, etc., with the permission of the University and respective supervisor*

# *Date : \_\_\_\_\_\_\_\_\_\_\_\_\_\_*

|  |  |
| --- | --- |
| ***Author(s):*** | |
| *Name: Mohammad AL Nawayseh* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  |
| *Name: Nehad Dhoun* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  |
| *Name: Salma Jedayah* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  |
| *Name: Salsabeel Khaleel* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |

***Supervisor(s):***

*Name: Mohammad Rasool Al-Jaafrah Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

# Anti-Plagiarism Declaration

This is to declare that the above publication produced under the supervision of *Mohammad Rasool Al-Jaafrah* having title “*Keystroke Behavioral Biometrics Continuous Authentication Approach Using ML Algorithms*” is the sole contribution of the author(s) and no part hereof has been reproduced illegally (cut and paste) which can be considered as Plagiarism. All referenced parts have been used to argue the idea and have been cited properly. I/We will be responsible and liable for any consequence if violation of this declaration is proven.

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
| ***Author(s):*** | |
| *Name: Mohammad AL Nawayseh* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  |
| *Name: Nehad Dhoun* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  |
| *Name: Salma Jedayah* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  |
| *Name: Salsabeel Khaleel* | *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |

# **ACKNOWLEDGMENTS**

In the name of Allah, the Most Gracious, the Most Merciful

All praise and greatness to Allah, Lord of the Two Worlds and may Allah's prayers and peace be upon our Prophet Muhammad and his companions.

We would like to express our deepest gratitude to our parents, who have supported us throughout our academic journey. Their love and guidance have been invaluable and we couldn't have done it without them. We would also like to thank our university doctors, who have provided us with valuable guidance and knowledge throughout our project. Their expertise and support have been instrumental in the successful completion of our graduation project.

Lastly, we would like to express our heartfelt thanks to our supervisor, Eng. Muhammad Al-Jafraah, for his unwavering support and guidance. His expertise, encouragement and mentorship have been essential in the completion of our graduation project.

We are deeply grateful for all the support, guidance and help received from all of you. We hope to make you proud with our future endeavors.

# **ABSTRACT**

Due to their use for storing personal information, communicating, and completing financial transactions, smartphones have become an essential element of daily life. Making sure smartphones are safe has become crucial due to our increasing reliance on them. Traditional authentication methods, including passwords and PINs, can be easily forgotten or vulnerable to attacks like  guessing attacks ,smudge attacks and shoulder surfing . Despite being quick, simple, and inexpensive, these methods are not impervious to attack. resulting in a requirement for more secure solutions. Behavioral biometrics, particularly keystroke dynamics, have emerged as a possible approach in this area.

Machine learning has the potential to increase the performance and accuracy of continuous authentication systems. This method will lay the foundation for the development of a more secure and user-friendly authentication solution designed exclusively for smartphones. By using the power of ML and leveraging keystroke dynamics, smartphone security may be successfully enhanced, sensitive data secured, and overall user experience improved.

In order to overcome the limitations of traditional authentication methods, this project intends to provide a continuous behavioral authentication model based on user keystroke activities. The proposed continuous behavioral authentication model is built on a set of highest importance 30 features only . A Random Forest (RF) classifier evaluates the concatenated features. The lowest Equal Error Rate (EER) of 0.02% is achieved with the use of features importance, and the best accuracy results are 99.97%. The results show that the model was superior to other approaches.

**Keywords**: Keystroke dynamics , Behavioral biometrics , Authentication , Machine learning

# **List of Abbreviations**

BB-MAS Behavioral Biometrics Multi-device and Multi-Activity data from Same users

ACC Accuracy

ANOVA Analysis Of Variance

KD Keystroke Dynamics

HT Hold Time

KRT Key Release Timestamp

KPT Key Press Timestamp

EER Equal Error Rate

DD Down-Down

DU Down-Up

UD Up-Down

UU Up-Up

FRR False Rejection Rate

FAR False Acceptance Rate

KNN k-Nearest Neighbors

ML Machine Learning

RF Random Forest

DT Decision Tree

NB Naive Bayes

# **List of Table**

# 

**Table 3.1** : Keystroke dynamics basic statistics ................................................................. 19

**Table 3.2** Keystroke Extracted Features ..............................................................................21 **Table 3.3**All Features ClassificationResults ………………………………………………24

**Table 3.4**Used Feature Selection methods............................................................................ 26

**Table 3.5**Feature Selection with Classifiers........................................................................... 28

**Table 3.6** : Importance Based Analysis Selected Features......................................................30

**Table3.7** KD Approach Evaluation Metrics of Multiple Classifiers .......................................35

# **List of Figures**

**Figure 3.1** Keystroke Dynamics Captured Timing Features...........................................20

**Figure 3.2**Keystroke Features Heat map...................................................................... 22

**Figure 3.3** Keystroke Features Importance ..................................................................23

**Figure 3.4** Keystroke authentication model design ..............................................31

**Figure 4.1**Confusion matrix.................................................................................... 36

**List of Contents**

[Intellectual Property Right Declaration i](#_Toc137719909)

[Anti-Plagiarism Declaration ii](#_Toc137719913)

[ACKNOWLEDGMENTS iii](#_Toc137719914)

[ABSTRACT iv](#_Toc137719915)

[List of Abbreviations v](#_Toc137719916)

[List of Table vi](#_Toc137719917)

[List of Figures vii](#_Toc137719918)

[CH1 : INTRODUCTION 1](#_Toc137719919)

[1.1 OVERVIEW 1](#_Toc137719920)

[1.2 Problem statement 1](#_Toc137719921)

[1.3 Motivation 2](#_Toc137719922)

[1.4 Project Objectives and goals 2](#_Toc137719923)

[1.5 Project Scope 2](#_Toc137719924)

[1.6 Report organization 3](#_Toc137719925)

[CH2: Background and Literature Review 4](#_Toc137719926)

[2.1 Authentication 4](#_Toc137719927)

[2.1.1 Authentication Importance 4](#_Toc137719928)

[2.1.2 Authentication criteria categorization 5](#_Toc137719929)

[2.1.3 Traditional Authentication methods 7](#_Toc137719930)

[2.2 Biometric and Biometric-Based Authentication 8](#_Toc137719931)

[2.3 keystroke 9](#_Toc137719932)

[2.3.1 Keystroke As a Behavioral Biometrics 10](#_Toc137719933)

[2.3.2 Keystroke Applications In Security Fields 10](#_Toc137719934)

[2.4 keystroke authentication 11](#_Toc137719935)

[2.4.1 Keystroke Dynamics Authentication 11](#_Toc137719936)

[2.4.2 Fixed Text vs. Free Text Keystroke Dynamics Authentication 13](#_Toc137719937)

[Chapter summary 16](#_Toc137719938)

[**CH3 :Proposed Model Methodology** 17](#_Toc137719939)

[3.1 Dataset 18](#_Toc137719940)

[3.3 Keystroke Features Extraction 19](#_Toc137719941)

[3.4 Keystroke Features Heat map 21](#_Toc137719942)

[3.5 Keystroke Features Importance 23](#_Toc137719943)

[3.6 Feature Selection and Classification Methods 24](#_Toc137719944)

[3.7 Keystroke Dynamics Authentication Approach 30](#_Toc137719945)

[CH 4 : Experiments Results and Discussion 34](#_Toc137719946)

[CH5: Conclusions and Future Work 37](#_Toc137719947)

[References 38](#_Toc137719948)

# 

# 

# **CH1 : INTRODUCTION**

## **OVERVIEW**

Smartphones have become a crucial part of daily life due to their role in storing personal information , communication and conducting financial transactions. With the growing reliance on smartphones, ensuring their safety has become important. Traditional authentication methods, such as passwords and PINs, are vulnerable to a variety of threats and easily compromised. As a consequence, more efficient and secure authentication systems are needed .

This project attempts to improve smartphone security by developing a continuous authentication model based on keystrokes dynamics . The unique patterns and traits that people display while typing on a smartphone keyboard are referred to as keystroke dynamics.

The main target of this project is to apply machine learning methods to create an effective authentication model that continuously monitors and analyzes the user's keystroke dynamics . By combining keystroke dynamics into the authentication process , it is possible to create a technique of confirming a user's identity that is more trustworthy and secure.

The drive for the project come from shortcomings in traditional authentication methods and the growing need to improve smartphone security . The overall security of smartphones can be improved by using machine learning algorithms and keystroke dynamic analysis, while providing a smooth and secure user authentication experience.

## **Problem statement**

With the widespread use of smartphones for personal and professional purposes, ensuring the security of user data has become a crucial concern. Traditional password-based authentication systems can be vulnerable to attacks such as phishing and brute force attacks. Also, Traditional biometric authentication methods have limitations such as false positives, false negatives, and spoofing attacks. Continuous authentication based on free text keystroke behavioral biometrics offers a more secure and reliable alternative. However, static keystroke patterns may not be enough to distinguish between genuine and fraudulent users. This project aims to develop a keystroke behavioral biometrics continuous authentication approach that addresses these limitations by analyzing dynamic keystroke patterns using machine learning algorithms. The proposed approach aims to provide a more secure and reliable authentication method for various applications, such as smartphone authentication, online transactions and accessing sensitive information.

## **Motivation**

The motivation for this project come from the need for improved authentication techniques for smartphones . Current authentication techniques are vulnerable to unauthorized access and lack real-time intrusion detection , creating threats to user privacy and security . Fortunately , a more robust and continuous authentication system for smartphones can be created by combining keystroke dynamic and ML , which offers improved security and user experience on smartphones.

## **Project Objectives and goals**

The main objective of this project is the implementation of a continuous authentication model based on free-text keystroke dynamics using Machine Learning Algorithms. The following objectives are taken into consideration:

1- Analysis of Keystroke dynamics data to identify the collection of features that will improve the proposed model's ability to identify unauthorized entry attempts.

2- Present and evaluate the effectiveness of Keystroke dynamics selected features through a unimodal using ML.

3- Conduct a thorough comparison of the proposed mode with others Machine Learning Algorithms regarding the goal of superiority demonstrating.

## **Project Scope**

The project scope involves the creation and deployment of an ML model for continuous authentication on smartphone utilizing keystrokes dynamics . It includes Using Latest Free text keystroke dataset, training the ML model with appropriate algorithms, and tackling the issues of continuous authentication. The objective of this project is to increase security, user experience , and overcome the limitations of traditional authentication techniques.

## **1.6 Report organization**

This project consists of five chapters organized as follows:

[CH1 : INTRODUCTION](#_Toc137422714)

CH2: Background and Literature Review

[CH3: Proposed Model Methodology](#_Toc137422734)

[CH 4 : Keystroke Dynamics Experiments Results and Discussion](#_Toc137422741)

[CH5: Conclusions and Future Work](#_Toc137422742)

# **CH2: Background and Literature Review**

The importance of authentication in digital information security is discussed in this chapter. Biometrics, specifically keystroke biometrics, are being investigated as a potential authentication technique. The notion of authentication will be looked at, as will the benefits of biometric-based authentication, keystroke applications as a behavioral biometric, and various methods of keystroke authentication, as well as the benefits and drawbacks.

## **2.1 Authentication**

The process of confirming a user's identity when they want to access a computer system or network is known as system authentication. It is essential to information security because it creates a relationship of trust between the user and the system [3]. It is crucial to make sure that only people with permission can access the large amount of sensitive data that is stored and transmitted digitally [1]. System authentication contributes to the prevention of illegal access, data breaches, and other security concerns. As a result, it is a critical component of any organization's IT security strategy. This essay will investigate the numerous forms of system authentication mechanisms and their importance in protecting the security of computer systems and networks [2].

Passwords, biometric factors such as fingerprint or facial recognition, and smart cards can all be used to authenticate users. Each method's level of security varies, with some ways being more secure than others [4]. Furthermore, multi-factor authentication, which employs two or more authentication methods, is gaining popularity due to its increased security [5].

### **2.1.1 Authentication Importance**

It is impossible to exaggerate the value of authentication. The risk of data breaches and other security concerns is rising as more and more data is being stored and exchanged digitally. By guaranteeing that only people with permission can access the data, authentication helps to reduce these dangers. It is also necessary to comply with a number of regulatory and industrial restrictions, which makes it a crucial component of every organization's IT security strategy. In the end, authentication is essential for safeguarding sensitive data and making sure that computer systems and networks are honest and private [6].

### **2.1.2 Authentication criteria categorization**

Authentication criteria categorization is a process of grouping authentication methods based on various criteria. The primary criteria for categorizing authentication methods are based on something the user knows, something the user has, or something the user is [7].

Typically, the criterion based on something the user knows are passwords or passphrases. To acquire access to the system or network via this method, users must supply a secret code that only they know.

Physical tokens such as smart cards or USB tokens are examples of criteria based on something the user has. These tokens give a unique identification that is linked to the user's account and is used to authenticate the user when they try to access the system or network.

Biometric authentication techniques like fingerprint or face recognition are among the criteria based on something the user is. When a user tries to enter the system or network using this approach, their distinctive physical attributes are used to authenticate them.

Other criteria for categorizing authentication techniques include the number of components necessary for authentication, the amount of security given, convenience of use, and cost [8].

Overall, categorization of authentication criteria is significant since it assists companies in selecting the most suitable authentication technique based on their individual needs and requirements. Organizations may guarantee that they are employing the most effective and efficient way for safeguarding their systems and networks by categorizing authentication methods [8].

Criteria for Authentication The level of security provided by each technique can also be used to categorize methods. Single-factor authentication techniques, such as passwords or passcodes, for example, are sometimes seen as less secure because they rely on only one factor to validate the user. Multi-factor authentication systems, on the other hand, which require two or more factors to authenticate the user, are deemed more secure since they add an extra layer of protection against illegal access [4].

The simplicity of use of various authentication techniques can also be used to classify them. Despite being relatively simple to use, some authentication mechanisms, like passwords, can be subject to phishing or brute-force assaults. Other approaches, such biometric authentication, can be more reliable but may also be trickier to use or call for specialist gear, like fingerprint scanners or cameras for facial recognition [4].

When classifying authentication techniques, cost is another crucial factor. Although the cost of implementing some techniques, like using passwords, is very low but may necessitate  additional resources, Other solutions, such as smart cards or biometric authentication, may be more costly to establish, but they may provide better security and require less ongoing maintenance [9].

Another criterion for categorizing authentication techniques is the number of factors necessary for authentication. Passwords and other single-factor authentication systems require only one factor to authenticate the user. Multi-factor authentication systems, on the other hand, necessitate the usage of two or more factors in order to authenticate the user. The factors can come from one of three sources: something the user knows, something the user has, or something the user is [10].

Because of its increased security, multi-factor authentication is becoming increasingly common. Two-factor authentication (2FA), for example, necessitates the user providing two factors, such as a password and a security token or a fingerprint scan. This adds another layer of security against unauthorized access, making it more difficult for attackers to obtain access to the system or network [11].

The level of security given is another key criterion for categorizing authentication systems. Some methods, such as passwords, are regarded as less secure because they are open to brute-force or phishing attempts. Other systems, such as biometric authentication or smart cards, are thought to be more secure since they are more difficult to copy or counterfeit [12].

Finally, when categorizing authentication systems, ease of use is a significant consideration. Some solutions, like as passwords, are simple to implement but difficult to remember or administer. Biometric authentication, for example, may be more convenient, but it may necessitate specialist hardware or training [8].

In conclusion, categorizing authentication requirements is a crucial procedure for enterprises to consider when picking an authentication technique. Organizations can select the most appropriate solution for their specific needs and requirements by understanding the various criteria and methods available.

### **2.1.3 Traditional Authentication methods**

In order to confirm a user's identity, traditional authentication techniques frequently rely on something the user know , like a password or passphrase. These techniques have a number of shortcomings despite their lengthy history of use and ease of use.

One of the key disadvantages of old authentication systems is that they might be exposed to attacks such as brute-force attacks or phishing attacks. Brute-force attacks include trying every conceivable combination of characters until the proper password is guessed, whereas phishing attacks involve fooling the victim into providing their password or other sensitive information [8].

Traditional authentication techniques can often be challenging to administer, especially in larger enterprises with a huge user base. Passwords should be secure and frequently changed, but they should also be challenging for users to remember. These problems can be addressed with password management software, but it can be costly and challenging to set up [4].

Last but not least, conventional authentication techniques only support single-factor authentication, which means they only use one factor to confirm the user's identity. When compared to multi-factor authentication, which involves two or more factors to authenticate the user, this may be less secure [4].

In summary, traditional authentication techniques have a number of problems, including vulnerability to attacks, difficulties in management, and absence of multi-factor authentication, even though they are simple to install. Because of this, a lot of firms are switching to newer, safer authentication techniques including biometric authentication, smart cards, and multi-factor authentication.

Traditional authentication techniques can be vulnerable to social engineering attacks in addition to the disadvantages already outlined. Social engineering attacks entail deceiving or manipulating the user into exposing their password or other sensitive information. For instance, a hacker may impersonate a reputable person or business and request the user to provide their password or other important information [13].

The need to remember numerous passwords for various accounts, which can be challenging and cause password fatigue, is another drawback of traditional authentication techniques . Users may as a result choose to use weak passwords that are simple to guess or the same password for numerous accounts, leaving them more open to attack [13].

Finally, attacks that take advantage of holes in the authentication procedure itself can be successful using conventional authentication methods. Attackers might, for instance, be able to intercept sensitive data such as passwords as it is being stored in memory or transferred over a network [4].

Many firms are switching to newer, more secure authentication techniques like biometric authentication, smart cards, and multi-factor authentication as a result of these and other downsides. For enterprises that need to safeguard sensitive data and systems, these techniques offer higher security and are more resistant to attacks.

## **2.2 Biometric and Biometric-Based Authentication**

Biometric authentication is a method of verifying an individual's identity based on their distinct physical or behavioral traits. It uses technology and science to measure and evaluate characteristics such as fingerprints, retinas, DNA, voice, and so on. Because these features are unique to each person, biometrics provides a dependable and secure form of identification and access control. As it includes measuring and analyzing characteristics of a person's body to authenticate their identification, the term "biometrics" comes from the Greek words "metric" meaning to measure and "bio" meaning life [14] .

Physiological and behavioral biometrics are the two main divisions. Physical characteristics that are distinctive to each individual and are challenging to modify include physiological biometrics like fingerprint, face, iris, and hand shape. A person's voice, signature, gait, and typing rhythm are examples of behavioral biometrics, which are tied to how they act or behave and can evolve over time as a result of psychological factors. While behavioral biometrics are regarded as "soft" biometrics and are well known for their dependability and accuracy in surveillance applications, physiological biometrics are typically more reliable [15].

In order to confirm someone's identification, biometric authentication requires collecting their biometric data and comparing it to a template that has been preserved. The identification of the person is verified if the biometric data matches the template that is saved [15].

However, biometrics have various advantages, including quick and secure authentication, ease, time savings, and versatility. However, there are concerns about the feasibility of certain biometric types for specific authentication scenarios, as well as the possibility of biometric security being compromised by hacking or theft of biometric data. Researchers are developing multi-biometric and hybrid biometric approaches to solve these challenges. Multi-biometrics is the use of various biometric features for authentication, such as fingerprints from separate fingers. In contrast, hybrid biometrics integrate several types of biometrics, such as iris and fingerprint, to improve security and accuracy. These methods seek to overcome constraints and improve the overall effectiveness of biometric authentication systems [14].

## **2.3 keystroke**

A keystroke, often called a key press, is the action of pressing a key on a keyboard or other input device. The computer receives a signal when a key is pressed on a keyboard, interprets the signal, and produces the intended result. A keyboard's keys each represent a certain character or function. Keystrokes, which are used to type letters, numbers, symbols, and other characters, are the basic text input units on a computer. They can also be used to enable keyboard shortcuts and other features in software programs [16].

Every keypress is accompanied by a unique scan code that identifies the key that was pressed. The scan code is acquired by the computer's CPU, which then uses the existing keyboard layout to translate it into a character or command [16].

Keystrokes can be recorded and analyzed for security purposes or to keep track of user behavior. For example, keystroke logging a technique for keeping track of and recording each keyboard operation made on a computer can be useful in spotting unauthorized access or other security problems [16].

### **2.3.1 Keystroke As a Behavioral Biometrics**

Keystroke dynamics, another name for typing biometrics, is a behavioral biometric system that verifies an individual's identity by using the particular characteristics of their typing habit [17].

Keystroke dynamics research began with desktop and laptop computers and has since expanded to mobile devices such as smartphones [18] .

Many features of a user's typing habit, including keystroke speed and rhythm, keypress duration, and time in between keystrokes, are determined by keystroke biometrics. These traits can be utilized to create a special "typing profile" for every individual, which can then be used to confirm that person's identity when typing on a keyboard [17].

In order to create a typing profile for a user, the system frequently gathers a sample of the user's typing activity (for instance, as they type a particular passage of text). This sample is then used to create a model of the user's typing behavior, which may subsequently be used to identify the user by comparing it to subsequent typing samples [17].

The fact that keystroke dynamics are easy to use and non-intrusive makes them more effective than other biometric authentication techniques. For instance, users can be confirmed by simply using a keyboard as normal [17].

### **2.3.2 Keystroke Applications In Security Fields**

One of the most modern security methods is keystroke dynamics technology. In this sector, it has numerous applications, including:

* Identification and verification. First, there is identification, which establishes a user's identity when no prior information about that person is available. This method compares a test sample against every template in a database. The system assigns the user the identification of the individual whose template most closely reflects the test sample. Authentication confirms the user's identity, which meets the second aim. The user provides his identify, and the system is responsible for confirming that the user is who he claims to be [19]. Free-text keystrokes system is used for both identification and authentication [19].
* Intrusion Detection: The free-text technique's continuous authentication system acts as a highly effective intrusion detection method. It is mostly used to detect any cautions about irregularities in a certain user's typing patterns. Free-text keystroke systems are also utilized for active system monitoring, which can aid in promptly and correctly identifying any breach. The creation of false alarms in ongoing keystroke-based authentication systems is a critical issue that must be handled in this situation. When they occur incorrectly, they may create frequent and sudden system halts that considerably upset users [19].
* Emotion Detection: Because free-text keystroke systems collect a lot of information from users during computer use, this information can also be utilized to infer the user's emotional state while they type. This has been used to determine how the user is feeling throughout the day while free typing. The typing activity of the user was used to infer emotions such as attention, anger, stress, relaxation, excitement, and tiredness [19].

## **2.4 keystroke authentication**

Keystroke authentication is a biometric-based authentication technique that identifies users by their distinctive typing patterns. In order to develop a distinctive typing profile, it is necessary to analyze a number of aspects of a person's typing activity, including time, pressure, rhythm, key sequences, and error rates. The individual's typing habits are then matched to the enrolled profile to confirm their identification, and this profile is utilized for authentication purposes [7] .

### **2.4.1 Keystroke Dynamics Authentication**

The temporal aspects of individual keystrokes can characterize an individual's typing rhythm, which is significantly influenced by their typing patterns and keyboard experience. This typing rhythm has a level of constancy that does not alter rapidly over a short period of time. The identity of a user can be continually verified without interfering with their usage of the system by recognizing keyboard patterns as they enter text. Keystrokes from users can be utilized to detect intruders or to determine their educational level [7].

Continuous authentication systems seek to detect attackers by monitoring users' activity in real time, without interfering with their usual usage of the system. Because of its unobtrusiveness, transparency, and accessibility, keystroke dynamics is one of the biometric modalities that is often used in such systems [7].

Using software that keeps track of the timing and length of keystrokes while the user types, keystroke dynamics can be seen. Collecting keystroke dynamics for continuous authentication systems is a cost-effective alternative to other biometric modalities like fingerprint and iris biometrics because it doesn't call for additional hardware [7].

It should be highlighted that keystroke dynamics could not be as unique and persistent as other biometric modalities, which could reduce the accuracy of user identification. Nevertheless, it continues to be a helpful tool in continuous authentication systems, particularly when combined with other biometric and non-biometric authentication techniques to boost security and accuracy [7].

The use of a fixed text for authentication, such as a password or PIN, has been the subject of numerous researches on keystroke dynamics. The objective of these studies is often to find distinctive keyboard patterns that can be used to authenticate people through timing and duration analysis of keystrokes [7].

However, there are drawbacks to utilizing a fixed text for authentication. It is unable to be used to fresh, fixed texts and only offers a partial approximation of an individual's keyboard rhythm. Additionally, fixed text authentication is susceptible to a variety of attacks, including replay attacks, in which a hacker records a user's keystrokes and then repeats them to enter the system [7].

Free-text input for keystroke dynamics authentication has been suggested by researchers as a way to get around these restrictions. Because it captures the normal variance in typing behavior that takes place when typing different words and phrases, free-text input enables a more thorough investigation of a person's keyboard rhythm. Greater security is also achieved because it is more challenging for attackers to replicate a person's typing style across many texts [7].

Overall, free-text input offers better accuracy and security for authentication applications while fixed text authentication is a helpful starting point for examining keystroke dynamics [7].

Keystroke dynamics are used to represent the current status of research on free text-based continuous authentication systems. Free text input offers a more complete picture of a user's keystroke patterns, but there hasn't been much in-depth study on how to use this information for continuous authentication [7].

In the majority of recent studies, learning the input mode and categorizing users based on their keystroke patterns has been accomplished using traditional machine learning techniques. The accuracy and efficacy of continuous authentication utilizing free text input may be improved, nevertheless, by applying more sophisticated approaches like deep learning algorithms and neural networks [7].

Additionally, more thorough research is still required to create models that can precisely capture and examine the dynamic patterns and patterns of keystroke behavior while employing free text input. We can design more effective continuous authentication systems that are less vulnerable to assaults and can provide users a greater level of security if we have a deeper understanding of these patterns [7].

### **2.4.2 Fixed Text vs. Free Text Keystroke Dynamics Authentication**

When it comes to keystroke dynamics, fixed text and free text input are the two main areas of research. They are frequently employed for various applications, and each field has advantages and disadvantages of its own [7].

Fixed text input, such as a password or PIN, uses a predetermined list of words or phrases for the user to type. This kind of input mode is frequently used for authentication since it enables the creation of a particular keystroke pattern for every user. Given that it just requires a small number of inputs, fixed text input is relatively simple to implement and analyze. However, it may be subject to attacks like replay attacks and is less indicative of the user's general typing pattern [7].

The term "free text input" on the other hand describes a system that enables users to enter any words or phrases they want. For continuous authentication purposes, this type of input mode can be used and offers a more complete depiction of the user's typing behavior. The analysis is more challenging due to the increased number of inputs, and it could be more challenging to create precise models for this kind of input method [7].

In general, both fixed text and free text input have benefits and disadvantages of their own, and they are frequently employed for various applications. Fixed text is frequently used for authentication, but free text is used for applications that need a more thorough awareness of the user's typing habits, such as continuous authentication [7].

* Fixed text input requires the user to repeatedly enter pre-defined text, such as a password or PIN. This information can then be utilized to build a model or train a system using statistical or machine learning methods. Using examples that are related to and unrelated to the user, the model can be labeled.

When the user needs to be recognized later, they will type the identical pre-defined text once more. This fresh sample will either be fed into the machine learning algorithm or linked to the previously stored sample. The algorithm may then assess whether the fresh sample matches the user's particular input pattern by comparing it to the keystroke data that was previously saved.

Since fixed text input enables the creation of a distinctive keystroke pattern for each user, it is frequently used for authentication. However, it may be subject to attacks like replay attacks and is less indicative of the user's general typing pattern.

Overall, fixed text input is a good place to start when examining keystroke dynamics and creating user authentication models. It has several limits, though, and more sophisticated methods, like free text input, might provide better accuracy and security for continuous authentication.

* Free text: in this instance, users can either type long blocks of text to approximate the idea of free text or, in other circumstances, they can type anything they want, whenever they want, without any limitations. It is the chosen algorithm's responsibility to use this input to extract the proper features and build a model for each user. Users can enter the same text or a different text when they need to be recognized later. It should be about a subject of their choosing. The chosen algorithm should determine whether these samples are real or not, and whether they are connected to the user-stated writings.

Since fixed text input enables the creation of a distinctive keystroke pattern for each user, it is frequently used for authentication. In order to enforce user access policies during the authentication process, this technique is frequently utilized. Once the user has been authenticated, it could be challenging to enforce user access regulations because no information about the user who is really using the system can be obtained. Contrarily, free text input offers a more thorough picture of the user's typing habits, which can be helpful for identifying certain keystroke patterns and enhancing the precision of continuous authentication. This technique can be used to continuously monitor how the user interacts with the system and to apply different restrictions if a change in the user's keyboard pattern is noticed. However, the drawback of this method is that it constantly tracks the user, wasting system resources and invading their privacy. Furthermore, since there might not be a constant text prompt for the user to enter during the authentication process when employing free text input, it might be more challenging to enforce user access controls. Overall, the application-specific trade-offs between accuracy, security, and privacy determine the best input technique for continuous authentication using keystroke dynamics. Free text and fixed text input methods are frequently used for many applications, although each has benefits and disadvantages of its own. When implementing continuous authentication utilizing keystroke dynamics, it is crucial to carefully analyze the potential privacy concerns and resource needs of each technique.

There are numerous difficulties related to continuous identity verification based on keystroke dynamics, which is still an active area of research [7].

One of the most difficult difficulties is modeling individual keyboard action using a statistically constrained model with fixed time characteristics. This is especially difficult with free-text keystrokes, which are highly changeable and adaptable. To overcome this issue, academics are investigating ways to build stable models of free-text keystrokes that can be utilized for permanent identity identification. This includes creating new techniques for feature extraction, data standardization, and modeling that can accurately capture each user's unique typing patterns.

Designing a practical and accurate technique for keystroke authentication continuously is another challenge. In order to accomplish this, one must have a thorough understanding of the intricate dynamics and patterns exhibited by keystroke behavior, as well as the capacity to track and examine these patterns continually in real-time. By making it possible to create models for continuous authentication that are more precise and effective, advanced machine learning techniques like deep learning algorithms and neural networks may provide a promising answer to this problem. All things considered, continuous identity verification based on keystroke dynamics has the potential to be an important tool for identifying and authenticating users in a variety of applications. To make it a workable and useful option for authentication and security purposes, however, there are a number of issues that must be solved. To address these difficulties and enhance the precision and dependability of continuous authentication using keyboard dynamics, ongoing research and development are required.

## **Chapter summary**

Verifying a user's or a device's identity in order to prevent illegal access is called authentication. Keystroke dynamics are a biometric-based authentication mechanism that has become popular because it overcomes the limitations of conventional approaches like passwords. Utilizing physical or behavioral traits, biometric-based authentication has drawbacks like implementation costs and privacy problems, but also benefits like accuracy and simplicity. Keystroke authentication is a biometric-based authentication technique that identifies users by their distinctive typing patterns. In order to develop a distinctive typing profile, it is necessary to analyze a number of aspects of a person's typing activity, including time, pressure, rhythm, key sequences, and error rates. The individual's typing habits are then matched to the enrolled profile to confirm their identification, and this profile is utilized for authentication purposes.

## **CH3 :Proposed Model Methodology**

Keystroke dynamics KD commonly include analyzing the time characteristics such as key press or release duration and their latencies.

In general, all KD evaluations entail:

* Selecting a number of participants to collect their typing data.
* Record all type of keystroke attempts.
* Preprocess data, and Extract features for training and testing purposes.
* Use a single portion of data for training.
* Evaluate the other portion of data by testing.

Many researches offered different choices in each of the aforementioned phases. However, in this study, different choices are made to evaluate the proposed approach in regard to these phases.

This section presents an overview of the algorithmic details for the proposed free-text keystroke authentication model and its primary phases. The proposed model will comprise multiple stages for constructing a continuous authentication model based on Keystroke Dynamics (KD). Typically, the first stage in such models is data collection. However, since a publicly available dataset is utilized in this research, the first stage will be data pre-processing. The second stage involves feature extraction, where the raw data is transformed into a more interpretable and efficient format. The third stage entails selecting appropriate features by eliminating irrelevant and unnecessary features for classification, thereby reducing the dimensionality of the dataset. Finally, the last stage involves testing and evaluating the feature selection methods using a Machine Learning (ML) classification model. These stages will be repeated for the SD authentication approach.

## **3.1 Dataset**

The BB-MAS (Behavioral Biometrics Multi-device and Multi-Activity data from Same users) dataset includes data from phones using the Android operating system provided by typing. For a period of three months, data from approximately 3.5 million keystroke operations was gathered from 117 individuals who have different typing abilities. The participants were given a phone with an app that required them to type two different sections of a fixed text. Additionally, they were presented with a series of ten questions that had to be answered using at least 50 letters or characters .The participants were told to enter the sample as naturally as possible, and if necessary, they were allowed to utilize backspaces or shift the cursor [7] .

**3.2 Keystroke Data Pre-Processing:**

In this project, the same method used in( [7] ) work is applied. The collected data was considered to be raw data, which needs to undergo pre-processing before it can be used in subsequent stages. Pre-processing is a crucial data mining approach that transforms raw data into a more comprehensible, usable, and efficient format. A dataset is a collection of data items, which are referred to as attributes or features in data science terminology. Pre-processing aims to improve dataset accuracy by removing incorrect or null values, improving consistency by deleting data discrepancies or duplication, and making the dataset more thorough by filling in any missing attributes. Lastly, smoothing the data makes it more useful and understandable. Data quality assessment, data cleaning, data transformation, and data reduction are all part of the pre-processing step.

The statistics of the keystroke data were evaluated to determine the average and standard deviation of the keystroke features, providing information about common behavioral traits associated with typing. Table ‎3.1 shows the statistics of KD from the dataset, with each participant typing an average of 9,400 keystrokes on their phone.

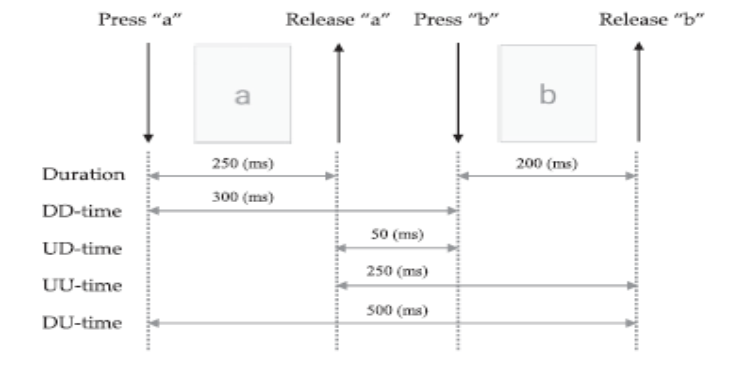
**Table 3.1** : Keystroke dynamics basic statistics

|  |  |  |  |
| --- | --- | --- | --- |
| Average | Standard Deviation | Min | Max |
| 9395 | **1472** | **5463** | **14694** |

During the analysis, a few outliers were identified in the timing data. Outliers are unusual data points that are three standard deviations above or below the mean and can be considered noise that could influence the overall performance of the system. To discover outlier features, a simple filter was applied to eliminate any instances of keys that were held down for two seconds or more. Additionally, any inter-key pauses lasting longer than a couple of seconds were removed, as they were believed to have been created by pauses in which the individual was either thinking or receiving instructions while collecting the data.

## **3.3 Keystroke Features Extraction**

Also, the feature extraction method proposed in ([7]) is applied in this project, Once the uni-graphs and diagraphs have been collected from the participants' raw data, the keystroke features are extracted. These features are computed using two significant values: the Key Press Time (KPT) and Key Release Time (KRT) in milliseconds for each key (n) and key-pair, as shown in Figure (3.1).



**Figure 3.1** Keystroke Dynamics Captured Timing Features

There are two main types of keystroke features that are extracted from the collected data:

1. Hold Time (HT): This is the time it takes for a key to be pressed and released. For each unigraph, the key hold time is extracted as a feature.

2. Keystroke Latencies: These are features that calculate the time elapsed between two actions made on two different keys. There are four types of latencies:

- Down-Down (DD): This is the time between two consecutive key presses.

- Up-Up (UU): This is the time between two consecutive key releases.

- Up-Down (UD): This is the time between key release and the next key press.

- Down-Up (DU): This is the time between pressing a key and releasing the next key.

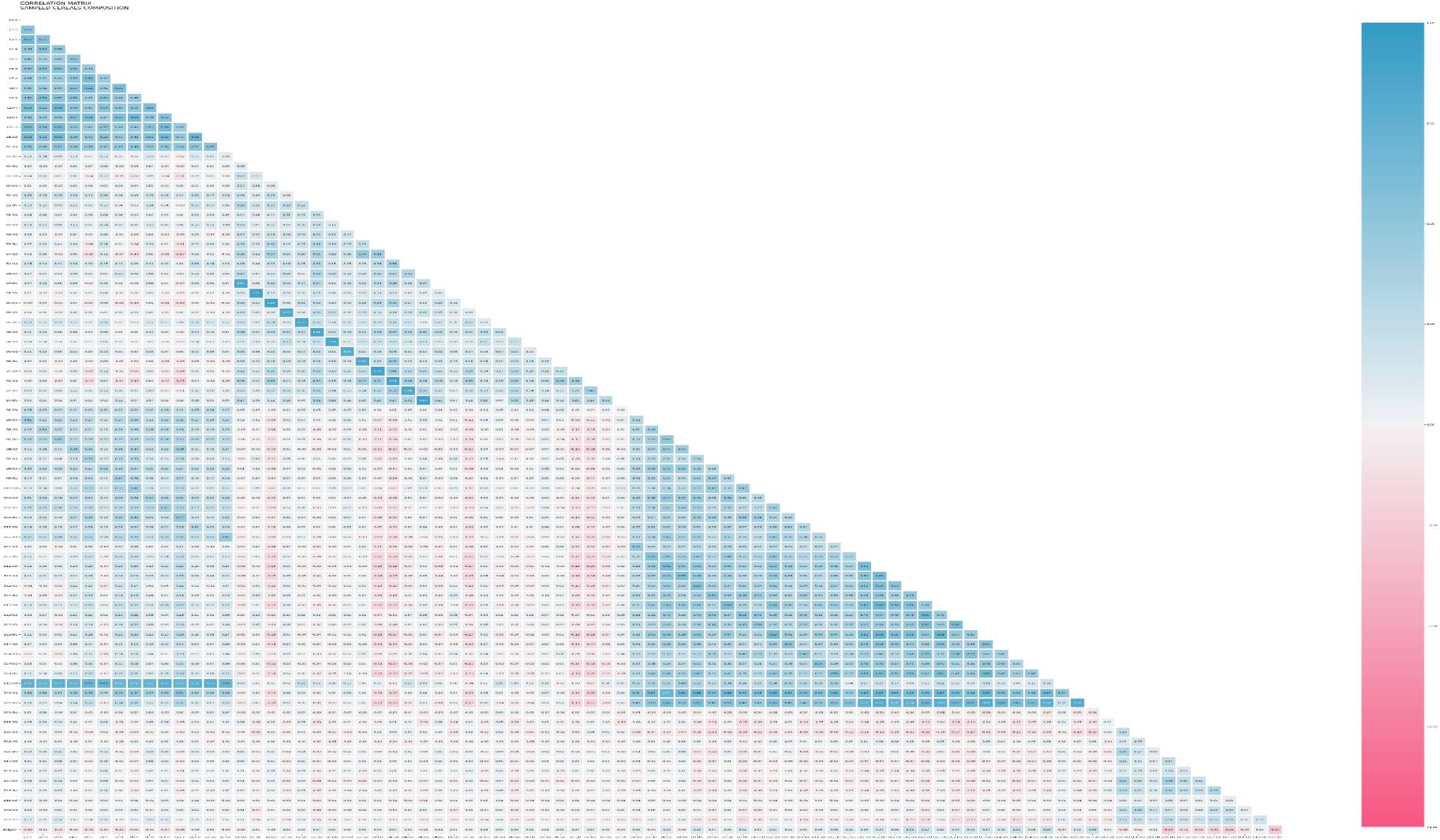
For each diagraph, four timing latencies features were defined, resulting in a total of 84 timing features extracted from all uni-graphs and diagraphs (18 diagraph types). Table (3.2 ) displays all the 84 extracted features.

**Table 3.2** Keystroke Extracted Features

|  |  |  |
| --- | --- | --- |
| Category | Feature Set | Number of features |
| Uni-graphs | {U1-H……………. U12-H} | 12 |
| Diagraphs | {D1-DD, D1-UU, D1-UD, D1-DU  D2-DD…………...……... D2-DU  .  .  .  .  D18-DD…………………. D18-DU} | 72 |

## **3.4 Keystroke Features Heat map**

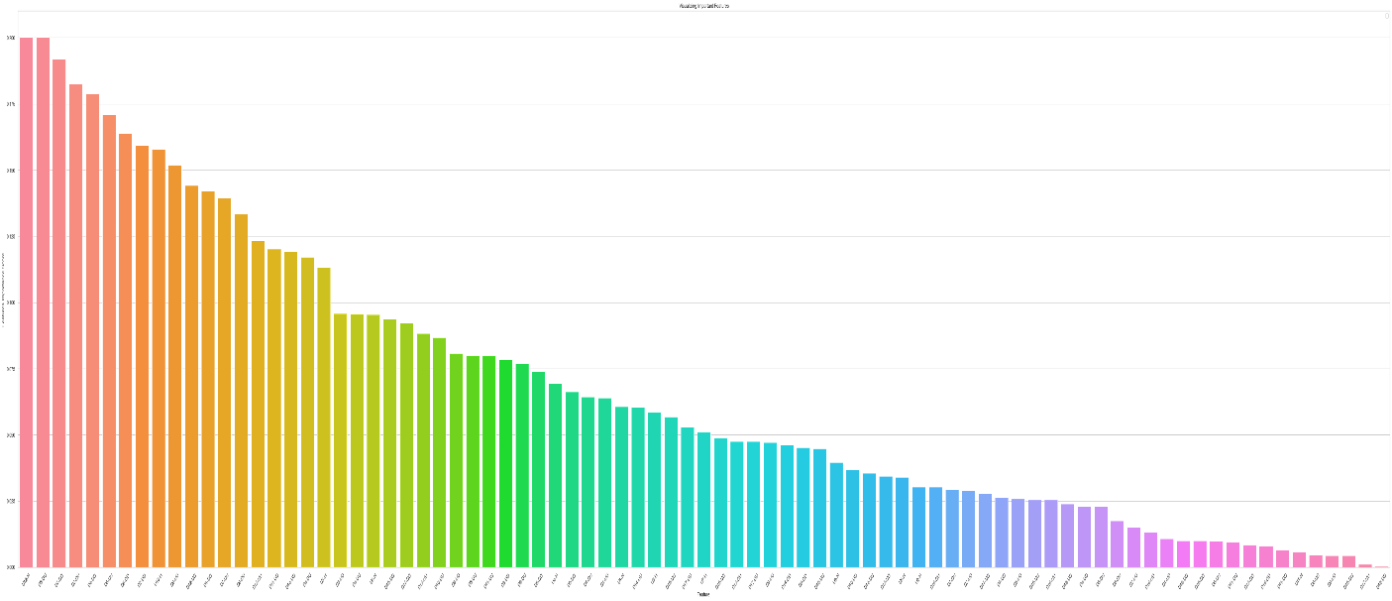
In order to show the density, intensity, patterns, variance, and anomalies of variables, heat maps are graphical tools that display data in a two-dimensional format using color maps. They depict variables on both axes to indicate relationships between them, and patterns can be seen by observing variations in color. Heat maps only accept numerical data and use various levels of color intensity to represent different values. The correlation coefficients between variables are displayed in correlation matrices, and a heat map grid can be used to visualize and pinpoint strong relationships. Strong dependence is shown by a positive correlation, whereas a strong inverse dependency is indicated by a negative correlation, and mild dependence is indicated by a correlation coefficient that is closer to zero. A correlation heatmap is used to examine the relationship between the qualities, as displayed in figure( 3.2) . This shows that a few features have a significant relationship whereas the remainder are not.



**Figure 3.2** Keystroke Features Heat map

## **3.5 Keystroke Features Importance**

The total feature relevance was investigated using the RF Classifier to better understand and interpret the heat map, and the findings are displayed in Figure (3.3). The KD features are ordered by decreasing significance for predicting the target class. A class of approaches known as feature importance assigns scores to input features in a predictive model, reflecting the relative importance of each feature when producing a prediction. These scores are beneficial in a variety of scenarios, such as improving data interpretation and minimizing the amount of input features in a predictive modeling challenge.



**Figure 3.3** Keystroke Features Importance

## **3.6 Feature Selection and Classification Methods**

This project addresses a multiclass classification problem in which samples are distributed to various user ID subjects without separating between positive and negative findings with 56 class labels. For multiclass classification, a variety of binary classification techniques can be utilized, such as modified binary classification algorithms that entail training different binary classification models for each class in comparison to all other classes or one model for each pair of classes. K-Nearest Neighbors (KNN), Decision Trees (DT), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM) are popular multiclass classification techniques.

To determine the optimum method for selecting features that best fit the keystroke data, all 84 features were applied to the previously selected classifiers and assessed using three metrics: overall accuracy, F score, Recall and classification precision. The results of applying all features to the specified classifiers are shown in Table (3.3 ).

**Table 3.3** All Features Classification Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | classifier | ACC | *F1 score* | *Recall* | *Precision* |
| 1 | **Random Forest** | **0.999** | **0.999** | **1** | **1** |
| 2 | K-Nearest Neighbors | 0.997 | 0.997 | 0.998 | 0.998 |
| 3 | XGBoost Classifier | 0.995 | 0.995 | 1 | 1 |
| 4 | Support Vector Machine | 0.989 | 0.989 | 0.993 | 0.993 |
| 5 | Multi-layer Perceptron | 0.989 | 0.989 | 1 | 1 |
| 6 | Naive Bayes Classifier | 0.969 | 0.969 | 0.976 | 0.976 |
| 7 | Logistic Regression | 0.962 | 0.962 | 0.97 | 0.97 |
| 8 | Decision Tree | 0.907 | 0.907 | 1 | 1 |

Overall, the findings illustrate that the Random Forest classifier achieved the highest overall accuracy, F1 score, recall, and precision when all 84 features were used. The Random Forest classifier achieved an accuracy of 0.999, an F1 score of 0.999, recall of 1, and precision of 1. This indicates that the Random Forest classifier was able to correctly classify almost all the instances, and the instances it classified as positive were almost always correct.

The other classifiers also performed well, with K-Nearest Neighbors achieving an accuracy of 0.997, XGBoost Classifier achieving an accuracy of 0.995, and the Multi-layer Perceptron achieving an accuracy of 0.989. However, the Support Vector Machine, Naive Bayes Classifier, Logistic Regression, and Decision Tree classifiers achieved lower accuracies, with the Decision Tree classifier achieving the lowest accuracy of 0.907.

In terms of F1 score, recall, and precision, the Random Forest classifier also achieved the highest scores, indicating that it was able to correctly identify positive instances with high accuracy and recall, while avoiding false positives with high precision. The other classifiers also achieved high scores but were slightly lower compared to the Random Forest classifier.

In summary, the results suggest that the Random Forest classifier is the best model for this task when all 84 features are used, achieving the highest overall accuracy, F1 score, recall, and precision. However, the other classifiers also achieved high scores, indicating that they may be suitable for the task depending on the specific requirements and constraints of the application.

Due to the relatively large number of keystroke features, there is a risk of dimension disaster, whereby the error increases with the number of features, and longer modeling time is required. Additionally, the inclusion of more features requires greater processing power to analyze the data and an increased amount of training data to generate meaningful models. In order to address this issue, a feature selection method was employed to achieve high accuracy rates using the fewest number of features possible. The objective of this process was to identify the most relevant features that contribute significantly to the performance of the model, while eliminating those that have little to no impact. The used feature selection methods are summarized in table (3.4 ) :

**Table 3.4** Used Feature Selection methods

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Type | Description | Reference |
| Random Forest Importance | multivariate | The tree node impurity is evaluated to determine whether it is pure or impure. Pure nodes consist only of observations from a single class, whereas impure nodes contain multiple observations from different classes. To evaluate the impurity of each node in a forest tree, it is measured both before and after a split is performed. This can be achieved using metrics such as the Gini index. The feature filter score Fk is calculated as the average reduction in impurity caused by Fk splits. A feature that is crucial for class prediction will produce a significant reduction in impurity on average. Therefore, such features are considered important for accurate classification and are selected for use in the model. | [7] |
| correlational analyses | multivariate | The Pearson correlation coefficient is a metric that measures the degree to which one variable is influenced by another. When two variables are highly correlated, they provide redundant information about the target. This means that we can accurately predict the target using just one of the correlated variables. Since the second variable does not provide any new information, removing it can help to reduce noise and dimensionality. By reducing the number of variables used in the analysis, we can simplify the model and improve its accuracy. This can be particularly important in cases where there are many variables to consider, and where data processing resources are limited. | [7] |
| Analysis Of Variance (ANOVA) | Univariate Statistical Tests | An analysis of variance (ANOVA) is performed on each feature in which the feature is represented by the class variable. The F statistic is calculated and used as the filter score. The F statistic measures how different the mean values of the relevant features are between the classes. The higher the F statistic, the more significant the difference in mean values of the corresponding feature between the classes. This means that features with higher F statistics are considered important for accurate classification and are selected for use in the model. By selecting only, the most important features, we can simplify the model and improve its accuracy. This approach is particularly useful in cases where there are many features to consider, and where data processing resources are limited. | [7] |

The created 84 feature subsets were applied to the feature selection methods described earlier. The methodology used was simple: the first feature was identified and tested, then the second feature was added to create a second group, and so on, until the number of features was increased cumulatively to achieve higher or equal results compared to applying all features to the same classification algorithm. This technique aimed to provide a comparative analysis of KD approaches to feature selection. The ANOVA, feature importance, and correlation-based selection techniques were compared using the best three classifiers, namely Random Forest, K-Nearest Neighbors, and XGBoost classifiers. Table (3.5) presents the best accuracy results, the number of features, and the classification time achieved by each approach.

**Table 3.5** Feature Selection with Classifiers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | ANOVA | IMPORTANCE | CORRELATION |
| Random Forest | **Accuracy %** | **99.04** | **99.97** | **99.97** |
| **Number of features** | **76** | **30** | **43** |
| **Time** | **1.14** | **2.11** | **11.45** |
| K-Nearest Neighbors | **Accuracy %** | **99.77** | **99.83** | **99.77** |
| **Number of features** | **76** | **30** | **43** |
| **Time** | **0.56** | **1.78** | **15.92** |
| XGBoost Classifier | **Accuracy %** | **99.21** | **99.89** | **99.75** |
| **Number of features** | **61** | **30** | **43** |
| **Time** | **490.31** | **4.98** | **12.60** |

The table shows the results of applying three different feature selection techniques (ANOVA, feature importance, and correlation-based) to three classifiers (Random Forest, K-Nearest Neighbors, and XGBoost) using keystroke dynamics data. The results compare the accuracy, number of features, and classification time achieved by each approach.

For the Random Forest classifier, the feature importance method achieved the highest accuracy rate of 99.97% using only 30 features and with a classification time of 2.11 seconds. The correlation-based method also performed well, achieving a 99.97% accuracy rate using 43 features and with a classification time of 11.45 seconds. The ANOVA method achieved the lowest accuracy rate of 99.04% using 76 features and with a classification time of 1.14 seconds.

For the K-Nearest Neighbors classifier, the feature importance and correlation-based methods achieved similar accuracy rates of 99.83% and 99.77%, respectively, using 30 and 43 features. The ANOVA method achieved a slightly lower accuracy rate of 99.77% using 76 features. The ANOVA method was the fastest with a classification time of 0.56 seconds, while the correlation-based method was the slowest with a time of 15.92 seconds.

For the XGBoost classifier, the feature importance method achieved the highest accuracy rate of 99.89% using 30 features, but with a longer classification time of 4.98 seconds. The correlation-based method achieved a slightly lower accuracy rate of 99.75% using 43 features and with a faster classification time of 12.60 seconds. The ANOVA method achieved the lowest accuracy rate of 99.21% using 61 features and the longest classification time of 490.31 seconds.

Based on these results, the feature importance method appears to be the most effective feature selection technique for optimizing the performance of the RF and KNN classifiers, achieving high accuracy rates with a relatively small number of features and reasonable classification times. For the XGBoost classifier, the correlation-based method could be a suitable alternative, achieving high accuracy rates with a slightly larger number of features and a faster classification time than the feature importance method.

In summary, the feature importance method is recommended for optimizing the performance of the RF and KNN classifiers based on keystroke dynamics data, while the correlation-based method may be a suitable alternative for the XGBoost classifier. The ANOVA method achieved lower accuracy rates with a larger number of features and reasonable classification times and may not be the optimal choice for feature selection in this context.

From a free-text keystroke dataset, this project evaluated the performance of various feature selection methods in conjunction with three classification methods, based on accuracy, classification time, and the number of selected features. The results indicate that no single feature selection method consistently outperforms all others. However, the feature importance method is recommended for optimizing the performance of the RF and KNN classifiers based on keystroke dynamics data, while the correlation-based method may be a suitable alternative for the XGBoost classifier. The ANOVA method achieved lower accuracy rates with a larger number of features and reasonable classification times and may not be the optimal choice for feature selection in this context.

To achieve the best result, the feature importance coefficients were calculated to rank the features, and then the Random Forest (RF) algorithm was used to apply these features. This approach achieved a 99.97% accuracy rate and reduced the number of features used by 64 %, while considering the memory and computing resources of the thin systems. Table (3.6) presents the features selected by the Importance-based analysis method. Overall, this method demonstrated that using high importance 30 features can enhance the accuracy of keystroke dynamics data for the RF classifier more efficiently.

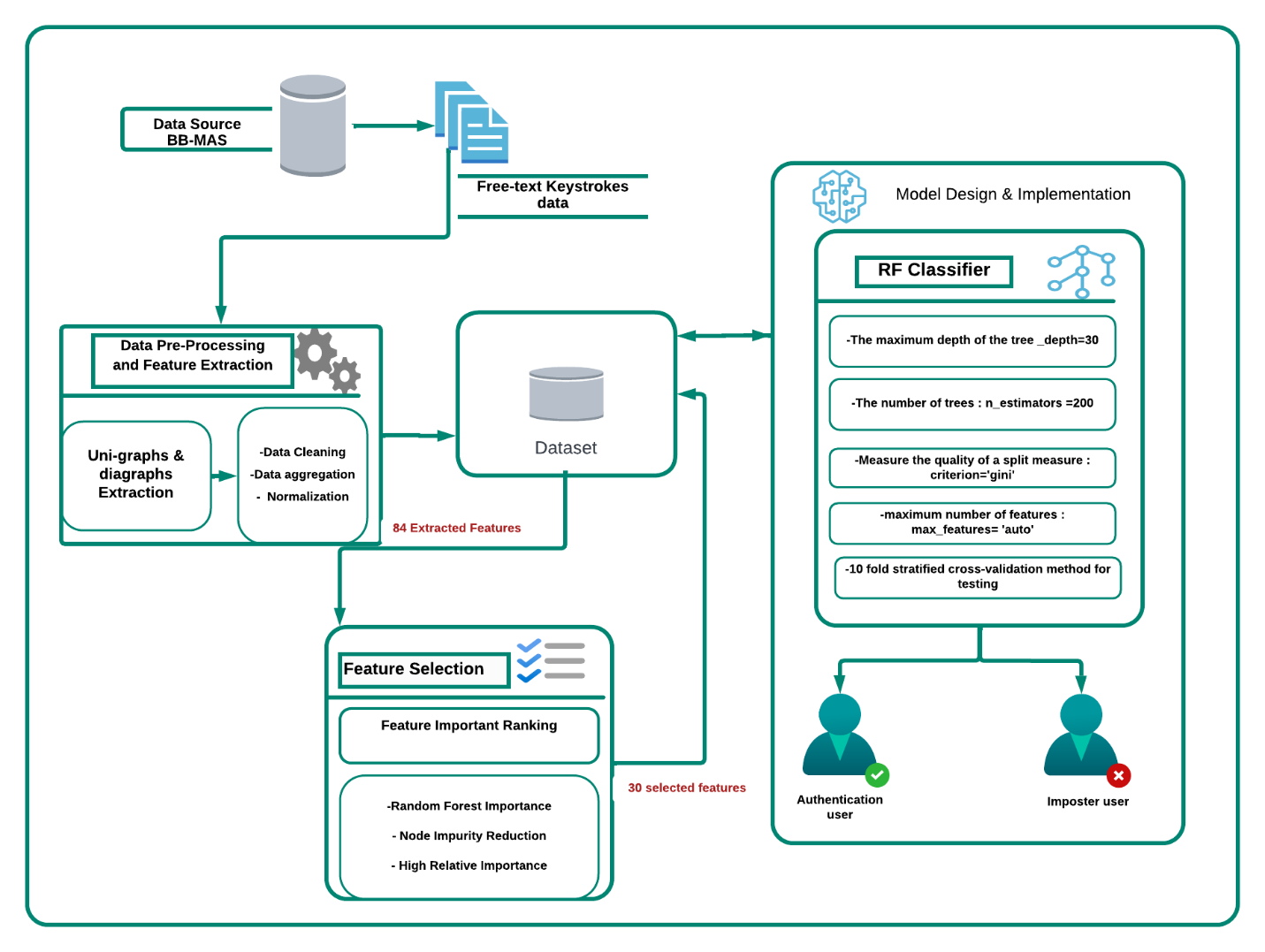
**Table 3.6** : Importance Based Analysis Selected Features

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Selected Features | | Number |
| Uni-graph | { “U11-H" , "U3-H"} | | 2 |
| Diagraph | **{"D15-DD","D15-UD","D15-UU","D6-UD","D3-UU","D10-D**  **D","D3-UD","D5-UU","D2-DD","D6-DU","D4-UD","D7-DD","D4-UU","D7-UD","D7-DU","D4-DD","D5-DD","D7-UU","D1-DU","D3-DU",D2-DU","D6-DD","D1-UD","D10-DU","D3-DD",D4-DU","D6-UU","D5-UD"}** | | 28 |
|  | | **Total** | **30** |

## **3.7** **Keystroke Dynamics Authentication Approach**

This section provides a detailed explanation of the Keystroke Dynamics (KD) model. Model selection involves the process of selecting the most appropriate model from a pool of candidate

models for a predictive task, using a variety of methods. Based on the comparison between classifiers and feature selection methods, it was determined that the Random Forest (RF) classifier, in combination with the feature Importance method was the most suitable approach for building a continuous authentication model based on free-text keystroke data. Figure (3.4) illustrates the design of the KD model.

 **Figure 3.4** Keystroke authentication model design

The KD model design, as shown in Figure ( 3.4 ), involved several steps. The first step was to gather and prepare the data by using BB-MAS dataset files to build a corpus of free text keystroke dynamics. Next, various data pre-processing and feature extraction techniques were implemented. This began with the extraction of the most frequent uni-graphs and diagraphs, followed by data cleaning processes such as missing value and outlier filtration. The cleaned KD data were then aggregated, and min-max normalization was applied. Overall, data pre-processing and feature extraction resulted in a total of 84 features.

The next step was to select the best subset of features using the importance-based feature selection method, resulting in a final set of 30 features. Finally, the data was loaded into the Keystroke Dynamics classification model, which was built using the RF classifier.

The Random Forest (RF) classifier has several hyperparameters that can be adjusted to improve the predictive power and efficiency of the model In this model, the following parameters were applied:

• The **n\_estimators** parameter, which specifies the number of trees constructed by the algorithm before averaging predictions or performing maximum voting. Increasing the number of trees improves performance and provides more stable predictions, but also slows down computation.

• The **max\_features** parameter, which determines the maximum number of features considered by the RF algorithm when splitting a node.

• The **max\_depth** parameter, which sets the maximum allowed levels in a decision tree. If set to None, the decision tree will continue splitting until purity is reached.

• The criterion parameter, which determines the function used to measure the quality of splits in a decision tree for classification problems.

• The **min\_samples\_leaf** parameter, which sets the minimum number of required leafs to split an internal node.

By adjusting these hyperparameters, the RF classifier can be optimized for the specific dataset and task at hand, improving the accuracy and efficiency of the Keystroke Dynamics model.

# **CH 4 :** **Experiments Results and Discussion**

This section describes the experiments conducted to compare the Keystroke Dynamic continuous authentication model with two other classifiers, namely, KNN and XGBoost. To ensure a fair comparison, the same number of features was used for each classifier. The feature selection and classifier methods were compared, and the random forest (RF) achieved the best results with the least number of features (30), using the importance-based feature selection method. Consequently, we used 30 features in this evaluation, which were applied to all other classifiers for comparison. The evaluation metrics used included the confusion matrix, accuracy, precision, recall, and F-measure.

Keystroke dynamics (KD) approaches are commonly evaluated using two core performance metrics: the false acceptance rate (FAR) and the false rejection rate (FRR). FAR is the ratio of identified unauthorized instances as authorized instances and falsely accepting them, while FRR is the rate of identified authorized instances as unauthorized instances and falsely rejecting them. Lower values for both metrics indicate a better performing system. In addition to FAR and FRR, two other performance metrics commonly used for evaluating systems are accuracy (ACC) and equal error rate (EER). ACC is a metric for evaluating classification algorithms and is calculated as the fraction of correct predictions made by the total number of predictions made. EER, also known as the misclassification rate, calculates the number of misclassified samples from both authorized and unauthorized classifications. The EER is the point where FAR and FRR are minimal and optimal, and lower EER values indicate higher accuracy of the biometric system. These metrics are used effectively to evaluate the performance of the Keystroke Dynamics model.

In this research, all metrics, measurements, and evaluation criteria were obtained from the PyCM version 3.1, which is a multi-class confusion matrix library, written in Python language.

With respect to the Keystroke Dynamics model, a comparison was performed to evaluate the effectiveness of all three algorithms. The same data set has been used with 30 features subset. Table (3.7 ) summarizes the results of multiple classification algorithms analysis obtained.

**Table3.7** KD Approach Evaluation Metrics of Multiple Classifiers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy  (%) | Precision  (%) | Recall  (%) | F-Measure  (%) | FAR  (%) | FRR  (%) | Error Rate  (%) |
| Random Forest\* | **99.97** | **99.97** | **99.96** | **99.30** | **0.00** | **0.00** | **0.02** |
| K-Nearest Neighbors | 99.84 | 99.82 | 99.83 | 95.47 | 4.49 | 0.8 | 1.06 |
| XGBoost | 99.89 | 99.87 | 99.86 | 96.84 | 3.15 | 0.6 | 0.14 |

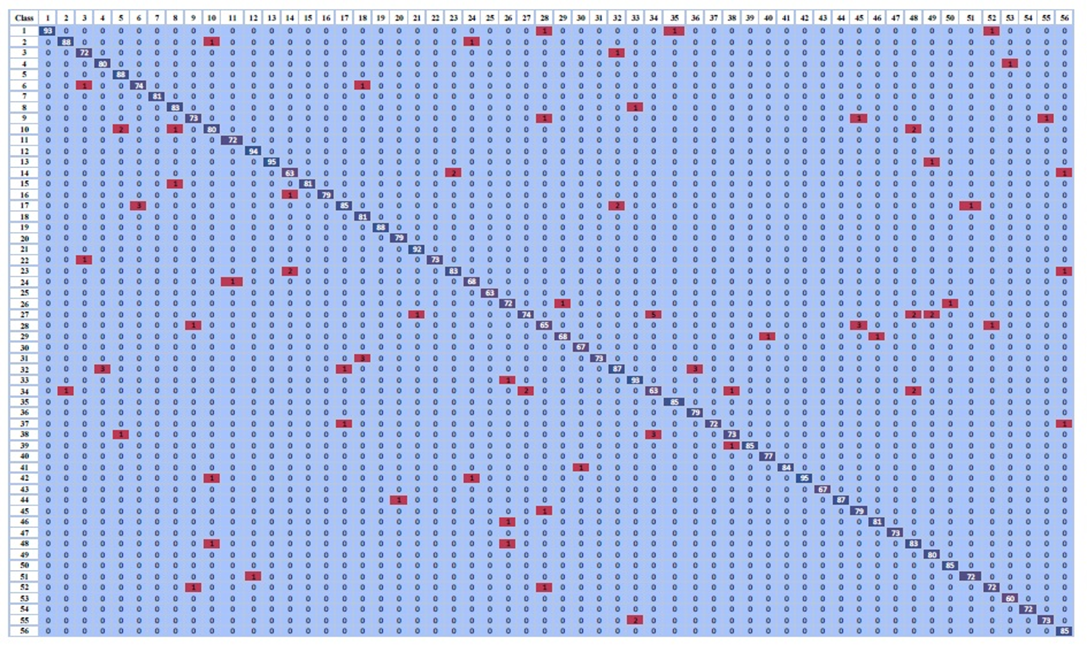
The results show that the Random Forest algorithm achieved the highest accuracy of 99.97%, as well as the highest precision and recall scores of 99.97% and 99.96%, respectively. However, its F-Measure score is lower than that of the K-Nearest Neighbors and XGBoost algorithms. The Random Forest algorithm also had a FAR and FRR of 0.00%, indicating that it had no false positive or false negative results.

The K-Nearest Neighbors algorithm had the highest F-Measure score of 95.47%, but it also had the highest FAR score of 4.49% and a relatively high FRR score of 0.8%. This means that it had a higher rate of falsely identifying unauthorized users as authorized and vice versa.

The XGBoost algorithm had the second-highest accuracy score of 99.89% and a lower FAR score of 3.15%, but it had a higher FRR score of 0.6% compared to the Random Forest algorithm.

Based on these results, the Random Forest algorithm appears to be the most effective classifier for the Keystroke Dynamics approach, as it achieved the highest accuracy, precision, and recall scores with no false positive or false negative results.

Looking at the confusion matrix, it becomes clear that the Keystroke Dynamics model achieved a high level of accuracy, correctly classifying most instances, with only a few examples being misclassified. The confusion matrix of the Keystroke Dynamics model is represented in figure (4.1).



**Figure 4.1**  Confusion matrix

# **CH5:** **Conclusions and Future Work**

Methods for improving smartphone continuous user authentication were introduced in this project . It was a uni-modal approach based on free-text keystroke dynamics.

The proposed uni-modal approach depends on using 56 users data from BBMAS dataset .

A Random Forest (RF) classifier with a feature-importance feature selection method was used to create the user authentication model based on keystroke dynamics. With only 30 features out of the 84 features in the dataset, exceptional performance was achieved with a detection accuracy of 99.97% and 0% for both FRR and FAR with the least EER. Based on the obtained data, our model proved to be superior when compared with other well-known ML classifiers and previous studies.

Future work will include implementing additional studies to obtain a fresh dataset by extending the participant pool and to cover more complex scenarios with many languages. Inquire as to whether deep learning techniques will help accuracy and other evaluation criteria even more.

# **References**

[1]​A. Mosenia, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, “CABA: Continuous Authentication Based on BioAura,” *IEEE Transactions on Computers*, vol. 66, no. 5, pp. 759–772, 2017, doi: 10.1109/TC.2016.2622262.

[2]​Y. Liang, S. Samtani, B. Guo, and Z. Yu, “Behavioral Biometrics for Continuous Authentication in the Internet-of-Things Era: An Artificial Intelligence Perspective,” *IEEE Internet Things J*, vol. 7, no. 9, pp. 9128–9143, 2020, doi: 10.1109/JIOT.2020.3004077.

[3]​K. W. Tse and K. Hung, “Framework for user behaviouralbiometric identification using a multimodal scheme with keystroke trajectory feature and recurrent neural network on a mobile platform,” *IET Biom*, vol. 11, no. 2, pp. 157–170, Mar. 2022, doi: 10.1049/bme2.12065.

[4]​G. Bhatt and B. Bhushan, “A comprehensive survey on various security authentication schemes for mobile touch screen,” in *Proceedings - 2020 IEEE 9th International Conference on Communication Systems and Network Technologies, CSNT 2020*, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 248–253. doi: 10.1109/CSNT48778.2020.9115731.

[5]​S. F. N. Sadikan, A. A. Ramli, and M. F. M. Fudzee, “A survey paper on keystroke dynamics authentication for current applications,” *AIP Conf Proc*, vol. 2173, no. November, 2019, doi: 10.1063/1.5133925.

[6]​S. J. Quraishi and S. S. Bedi, “Keystroke dynamics biometrics, a tool for user authentication-review,” *Proceedings of the 2018 International Conference on System Modeling and Advancement in Research Trends, SMART 2018*, pp. 248–254, 2018, doi: 10.1109/SYSMART.2018.8746932.

[7]​J. Al-Saraireh and M. R. AlJa’afreh, “Keystroke and swipe biometrics fusion to enhance smartphones authentication,” *Comput Secur*, vol. 125, p. 103022, Feb. 2023, doi: 10.1016/J.COSE.2022.103022.

[8]​I. Stylios, S. Kokolakis, O. Thanou, and S. Chatzis, “Behavioral biometrics & continuous user authentication on mobile devices: A survey,” *Information Fusion*, vol. 66, no. July 2020, pp. 76–99, 2021, doi: 10.1016/j.inffus.2020.08.021.

[9]​I. Stylios, S. Kokolakis, O. Thanou, and S. Chatzis, “Behavioral biometrics & continuous user authentication on mobile devices: A survey,” *Information Fusion*, vol. 66, no. August, pp. 76–99, 2021, doi: 10.1016/j.inffus.2020.08.021.

[10]​D. Cilia and F. Inguanez, “Multi-model authentication using keystroke dynamics for smartphones,” *IEEE International Conference on Consumer Electronics - Berlin, ICCE-Berlin*, vol. 2018-Septe, pp. 1–6, 2018, doi: 10.1109/ICCE-Berlin.2018.8576226.

[11]​E. M. Biometrics, R. Guest, and E. Ellavarason, “Multi-modal Verification,” no. 675087, 2021.

[12]​R. S. Choras, “Multimodal Biometrics for Person Authentication,” *Security and Privacy From a Legal, Ethical, and Technical Perspective*, 2020, doi: 10.5772/intechopen.85003.

[13]​M. Sharif, M. Raza, J. H. Shah, M. Yasmin, and S. L. Fernandes, “An overview of biometrics methods,” *Handbook of Multimedia Information Security: Techniques and Applications*, pp. 15–35, 2019, doi: 10.1007/978-3-030-15887-3\_2.

[14] M. Sharif, M. Raza, J. H. Shah, M. Yasmin, and S. L. Fernandes, “An overview of biometrics methods,” in *Handbook of Multimedia Information Security: Techniques and Applications*, Springer International Publishing, 2019, pp. 15–35. doi: 10.1007/978-3-030-15887-3\_2.

[15] I. Majeed Alsaadi, “Study On Most Popular Behavioral Biometrics, Advantages, Disadvantages And Recent Applications : A Review,” *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 10, p. 1, 2021, [Online]. Available: www.ijstr.org

[16] C. A. Iglesias, J. I. Moreno, D. Rivera, Institute of Electrical and Electronics Engineers, M. IEEE Systems, and Universidad Politécnica de Madrid. ETSI Telecomunicación‏, *Proceedings of the 2020 16th International Conference on Intelligent Environments (IE) : Madrid, Spain, 20-23 July*.

[17] P. H. Pisani and A. C. Lorena, “A systematic review on keystroke dynamics,” *Journal of the Brazilian Computer Society*, vol. 19, no. 4. Springer London, pp. 573–587, Nov. 01, 2013. doi: 10.1007/s13173-013-0117-7.

[18] G. Gibson, USENIX Association, Calif. USENIX Conference on File and Storage Technologies 4 2005.12.13-16 San Francisco, and Calif. FAST 4 2005.12.13-16 San Francisco, *FAST ’05 4th USENIX Conference on File and Storage Technologies ; Dec. 13-16, 2005, San Francisco, CA, USA*. USENIX Association, 2005.

[19] A. Sanghi, “Survey, Applications and Security of Keystroke Dynamics for User Authentication,” 2017.