




Article

# IKDD: A Keystroke Dynamics Dataset for User Classification

Ioannis Tsimperidis \* , Olga-Dimitra Asvesta, Eleni Vrochidou  and George A. Papakostas \* 

MLV Research Group, Department of Informatics, Democritus University of Thrace, 65404 Kavala, Greece; olasves@cs.ihu.gr (O.-D.A.); evrochid@cs.duth.gr (E.V.)

\* Correspondence: itsimper@cs.duth.gr (I.T.); gpapak@cs.duth.gr (G.A.P.)

**Abstract:** Keystroke dynamics is the field of computer science that exploits data derived from the way users type. It has been used in authentication systems, in the identification of user characteristics for forensic or commercial purposes, and to identify the physical and mental state of users for purposes that serve human–computer interaction. Studies of keystroke dynamics have used datasets created from volunteers recording fixed-text typing or free-text typing. Unfortunately, there are not enough keystroke dynamics datasets available on the Internet, especially from the free-text category, because they contain sensitive and personal information from the volunteers. In this work, a free-text dataset is presented, which consists of 533 logfiles, each of which contains data from 3500 keystrokes, coming from 164 volunteers. Specifically, the software developed to record user typing is described, the demographics of the volunteers who participated are given, the structure of the dataset is analyzed, and the experiments performed on the dataset justify its utility.

**Keywords:** keystroke dynamics; data mining; user classification; free-text dataset; biometrics



**Citation:** Tsimperidis, I.; Asvesta, O.-D.; Vrochidou, E.; Papakostas, G.A. IKDD: A Keystroke Dynamics Dataset for User Classification. *Information* **2024**, *15*, 511. <https://doi.org/10.3390/info15090511>

Academic Editor: Antonio Jiménez-Martín

Received: 22 July 2024

Revised: 12 August 2024

Accepted: 20 August 2024

Published: 23 August 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Keystroke dynamics is defined as the collection of biometric information related to events on the physical or virtual keyboard as a user types and then exploits it to derive various conclusions. The ability to identify someone by the way they use keys became apparent in the 19th century when telegraph operators could tell who was transmitting a message by tapping style [1]. The first research in keystroke dynamics, which aimed to authenticate users by recognizing each person’s unique typing pattern, were based on this observation. Since then, many studies have been published, in which various authentication techniques have been proposed, with very good results [2].

User authentication has received the most research attention in keystroke dynamics. However, many researchers approached the subject from a different angle.

Buker et al. [3] attempted to identify the user’s gender. They studied keystroke dynamics in live-chat interfaces on popular applications, such as WhatsApp and Skype. The results show that this feature is recognizable, with an over 95% success rate, and that there is a general tendency for female and male users to type differently, especially concerning the communication’s immediacy and the social aspects of the interaction.

Pentel [4] tried to predict the age of a user through their keystroke dynamics. Upon using binary classification, the best achieved f-score is over 0.92 and the worst is 0.82, while multiclass classification was able to sort all groups with an over the baseline accuracy.

Monaro et al. [5] tried to differentiate between true and false personal information that users type on a computer keyboard. The conclusions indicate that this method is able to distinguish the truths and lies in specific types of autobiographical information, with an accuracy higher than 75%.

Epp et al. [6] experimented with keystroke dynamics in order to figure out the emotional state of a user. In total, 15 classifiers for emotions were made. The emotions of confidence, hesitation, anxiety, tranquility, sadness, and tiredness granted the best results, with accuracy fluctuating between 77% and 88%.

Marrone and Sansone [7] focused on the use of keystroke dynamics as a way to continuously predict users' emotional states during message writing sessions. The conclusions differ according to the method used to process data. The highest accuracy is achieved using the multiple-instance learning–support vector machine (MIL-SVM) model when trained on variable-sized bags. Neutral and happy are the best recognized emotional states.

Kořakowska and Lndowska [8] analyzed keystroke dynamics while participants were writing positive and negative opinions. A support vector machine (SVM) model was used for classification. The best achieved F1-score was 0.76.

The standard features that are used in keystroke dynamics research are inter-key latency and key hold-time [9]. The former is a measure of the amount of time between a key being released and the subsequent key being pressed. The latter is a measure of the amount of time between a key being pressed and the same key being released. Inter-key latency can be further implemented as digram latency and n-gram latency. Digram latency is defined as the time needed by a user to use two consecutive keys, and similarly, n-gram latency describes the time needed to use n consecutive keys [10]. Digram latency can be expressed in four different ways: (1) the time needed from the press of the first key until the press of the next key (down–down digram latency, DDDL), (2) the time needed from the release of the first key until the release of the next key (up–up latency, UUDL), (3) the time needed from the press of the first key until the release of the next key (down–up latency, DUDL), and (4) the time needed from the release of the first key until the press of the next key (up–down latency, UDDL).

Pauses in keystroke dynamics indicate an unusual amount of time needed to press two keys subsequently. These intervals may occur either between words, different logical units of a text or due to external factors.

Typing speed can be considered as a measure that shows how experienced a user is regarding the usage of a computer keyboard. It is important to note that even though typing speed may be such an indicator, speed is not an important factor in keystroke dynamics.

All keystroke dynamics research, regardless of what their research goals are, needs data derived from user keystroke recording. Such datasets are rare in the literature and for the development of this research field it is imperative to have other keystroke dynamics datasets available. The main objective of the present work is to contribute to solving this problem by presenting such a dataset, whose comparative advantages over the others available are the following: (1) it contains data from the recording of users during their daily use of the computer, (2) it contains a large amount of data, with approximately 1.85 million keystrokes recorded, (3) it contains data that is characterized by five tags, namely gender, age group, handedness, mother tongue, and educational level of the users.

The rest of the paper is as follows: Section 2 reviews the available keystroke dynamics datasets. Section 3 describes the creation process of the keystroke dynamics dataset named IKDD, as well as its structure. In Section 4, some examples are given for the use of the dataset. Finally, Section 5 discusses and Section 6 summarizes the paper.

## 2. Keystroke Dynamics Datasets

The data used in the keystroke dynamics studies came from recordings of volunteers typing, either in fixed-text mode or in free-text mode. In the first approach, volunteers are asked to type a specific text, usually in the environment of an application. In the second approach, volunteers type whatever they want. The advantage of collecting data in fixed-text mode is that almost the same amount of data is obtained from each volunteer and the same keystroke dynamics features can be extracted from each logfile. The advantage of collecting data in free-text mode is that it better approximates the volunteers' actual typing and therefore the data collected are more representative.

One of the big problems of keystroke dynamics is that there are not many datasets available on the Internet, especially free-text datasets. The reason is that they may contain sensitive and/or personal information of the volunteers, such as passwords, credit card

numbers, messages to third parties of private interest, etc. The few available keystroke dynamics datasets have been used by a number of studies.

Killourhy and Maxion's dataset [11] comes from fixed-text, and more specifically from the typing recording of the password ".tie5Roanl". The data were collected from 51 subjects, 21 females and 30 males in particular, during eight data collection sessions. There were eight left-handed people, while the rest were right-handed. The median age group was 31–40 years and the youngest was 18–20 years and the oldest 61–70 years. Each session lasted between 1.25 and 11 min, with the median session taking about 3 min.

The "Keystroke100" dataset from Loy et al. [12] was developed from fixed-text, in the form of the password "try4-mbs". There were 100 participants in this study. Gathered data were collected from modified keyboards that measured not only duration but also pressure applied on each key being used.

The "KeyRecs" dataset, from Dias et al. [13], consists of both fixed-text and free-text samples. The dataset under consideration comprises about 1.6 million keystrokes gathered from 99 participants. Out of them, only 39 of the participants were female and the remaining were male, while only 8 were left-handed and the rest were right-handed. The study was performed mostly by individuals between the ages of 18 and 20, while the age range was 18–51 years. The participants were of 20 different nationalities, with the most common being from Poland, Portugal, Greece, and Italy.

Risto and Graven's dataset [14] was obtained with six passwords used as fixed-text. The collected data come from 103 participants, each of which typed two passwords ten times each, corresponding to roughly 3 participants to each password. The participants are predominantly university students with high variance in typing proficiency.

The "GREYC" dataset by Giot et al. [15] comes from fixed-text inputs. It involves 133 users, with 100 of them contributing samples from no less than five distinct sessions. The interval between each session was at least one week. Each user typed the password "greyc laboratory" 12 times, on two distinct keyboards during each session. There was also a possibility to change the given password, with the default one being "greyc laboratory". Each user is able to type in different passwords and a model is created for each one of them. The acquisition process took place between 18 March 2009 and 5 July 2009. Out of 100 participants, 32 were females and the rest were male, while the age range was between 18 and over 50 years, with most participants belonging to the 18–25 age group.

Sznur and García's dataset, "KEasyLogger" [16], consists of free-text data. This is the largest public keystroke-labeled dataset available to date. It comes from 17 individuals and the contained data were collected during 379 sessions.

Maalej and Kallel's "EmoSury" [17] dataset combines both free-text and fixed-text data. This dataset comes from a dynamic web application and was designed in the context of an experiment, aimed to understand a user's emotional state from their keystroke dynamics.

Clarkson University Keystroke Dataset by Vural et al. [18] includes keystroke data for short pass-phrases, fixed-text, and free-text. The data collection was conducted with a total of 39 subjects spanning a period of eleven months between August 2011 and June 2012. Each subject attended two sessions of approximately one hour each, on two separate days. This dataset also contains video of a subject's facial expressions and hand movements during data collection sessions.

El-Abed et al.'s [19] dataset "RHU" was obtained by typing the password "rhu.university". The data were acquired from 53 individuals who participated in the acquisition process. All participants participated in three sessions, giving a total of 985 acquisitions, about 17 from each user. Out of all the participants, 24 were females and the rest were males, while the age ranging from 7 to 65 years, with the most common age group being 19–29.

Table 1 depicts aggregately all the data of each dataset.

Something that seems to be missing from the literature is a keystroke dynamics dataset that comes from recording users during their daily use of their computers. Such a dataset will approximate in the best possible way the actual typing conditions and therefore the

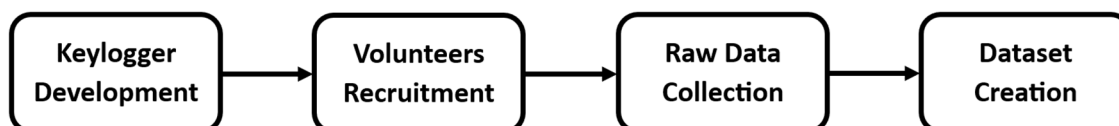
results extracted from its use will be more reliable. The IKDD, proposed in this paper, is a dataset created in this way and fills this identified gap in the literature.

**Table 1.** Datasets' characteristics.

Dataset	Number of Participants	Type of Recorded Text	Demographics	Software Used	Acquisition Period	Amount of Acquired Data
Killourhy and Maxion	51	Fixed	21 Females–30 Males 8 Left-Handed–43 Right-Handed 18–70 Age Range	Windows Application	N/A	400 Acquisitions
Keystroke100	100	Fixed	N/A	Developed Program	N/A	N/A
KeyRecs	99	Fixed and Free	39 Females–60 Males 8 Left-Handed–91 Right-Handed 18–51 Age Range 20 Nationalities	Online Platform	N/A	1.6 Million Keystrokes
Risto and Graven	103	Fixed	Mostly University Students	Keylogger	N/A	N/A
GREYC	133	Fixed	32 Females–68 Males 18–50+ Age Range	GREYC-Keystroke Application	18 March 2009–5 July 2009	7555 Acquisitions
KEasyLogger	17	Free	N/A	KEasyLogger Application	N/A	379 Acquisitions
EmoSury	N/A	Fixed and Free	N/A	Dynamic Web Application	N/A	N/A
Clarkson University Keystroke	39	Fixed and Free	N/A	JavaScript Keylogger	August 2011–June 2012	N/A
RHU	53	Fixed	24 Females–29 Males 7–65 Age Range	Touch Mobile Phones Application	N/A	985 Acquisitions

### 3. The IKDD Dataset

To create the IKDD, the following steps were completed: (1) A free-text keylogger was designed and implemented, (2) volunteers were recruited who agreed to participate in recording their daily keyboard usage, (3) the keystroke logging data were collected, (4) and specific keystroke dynamics features were extracted, which constitute the dataset. This process is shown in Figure 1.



**Figure 1.** Dataset creation flowchart.

#### 3.1. The Keylogger IRecU

For the needs of recording user typing, a keylogger named IRecU was designed and implemented. IRecU can only be run on devices with an MS Windows operating system. The first time the volunteer uses it, a window appears in which they state their demographic information, as well as a username, with which they will be identified each time they use the keylogger. When IRecU is running, an indication appears on the screen informing the user that it is being recorded.

IRecU was published in eight languages, specifically in Albanian, Arabic, Bulgarian, English, German, Greek, Malayalam, and Turkish, with the aim of creating a friendly environment for users whose mother tongue is one of these languages and therefore facilitating its use. IRecU runs only when the user wants it to and creates txt logfiles containing the recorded typing data. Each logfile contains data from approximately 3500 keystrokes.

### 3.2. Recruitment of Volunteers

For data collection, hundreds of people were approached to participate in the typing recording process. Of these, over 200 people accepted to participate, but 164 were able to successfully complete the process.

The process of recruiting volunteers and recording typing took place over three periods of time. Specifically, from 20 February 2014 to 27 December 2014, from 24 October 2017 to 28 May 2018, and from 29 March 2022 to 17 October 2023.

The volunteers were given instructions on the correct use of IRecU, as well as on the process of submitting the logfiles. Also, a consent form was signed, in which the possible risks of recording typing data were mentioned, while explicit commitments were given on the part of the researchers not to share the data with third parties and to use it exclusively for research purposes.

The demographics of the volunteers who submitted logfiles are presented in Table 2.

**Table 2.** Volunteer demographics.

Characteristic	Class	Number of Volunteers
Gender	Female	88
	Male	76
Age Group	18–25	49
	26–35	44
	36–45	40
	46+	31
Handedness	Left-Handed	14
	Right-Handed	146
	Ambidextrous	4
Mother Tongue	Albanian	17
	Bulgarian	18
	English	15
	Greek	106
Educational Level	Turkish	8
	ISCED-3	33
	ISCED-4	7
	ISCED-5	33
	ISCED-6	51
	ISCED-7–8	40

Regarding the educational level of the volunteers, in Table 2, due to the fact that data were collected from users living in different countries, and due to the fact that each country has its own educational system, it was decided to classify the level according to the International Standard Classification of Education (ISCED). Each educational level, of each educational system, is assigned to one of the nine ISCED levels (ISCED-0 to ISCED-8). For example, the primary education diploma is assigned to ISCED-1, while the doctoral degree is assigned to ISCED-8.

The volunteers were asked to use IRecU as much as needed to create three logfiles each. However, some of the volunteers decided to withdraw from the procedure before its completion, as was also provided as a possibility for the volunteers in the consent form, resulting in less than three logfiles being submitted. Also, some other volunteers continued their recording even after the end of the process, thus submitting more than three logfiles.

In total, users submitted 533 logfiles whose demographics are presented in Table 3.

**Table 3.** Logfiles' demographics.

Characteristic	Class	Number of Logfiles
Gender	Female	279
	Male	254
Age Group	18–25	151
	26–35	149
	36–45	125
	46+	108
Handedness	Left-Handed	50
	Right-Handed	471
	Ambidextrous	12
Mother Tongue	Albanian	51
	Bulgarian	54
	English	58
	Greek	345
	Turkish	25
Educational Level	ISCED-3	100
	ISCED-4	23
	ISCED-5	110
	ISCED-6	173
	ISCED-7–8	127

As can be seen from Table 2, the female volunteers are almost equal in number to the male volunteers, as is approximately the case in the world population. Also, right-handed volunteers are about 90% of all volunteers, as predicted by studies [20]. The volunteers are almost evenly distributed both in the age group they belong to and in their level of education, with the only exception being the ISCED-4 educational level. The reason for this is that the ISCED-4 educational level is not included in many educational systems, while in those that are included, such as post-secondary non-tertiary education, it does not have many graduates, compared to the other educational levels. Similar observations are made in Table 3.

### 3.3. Format of Raw Data

The logfiles created by IRecU consist of records and have the following format:

```
32,#2022-03-29#,72698762,"dn"
32,#2022-03-29#,72698856,"up"
90,#2022-03-29#,72699012,"dn"
65,#2022-03-29#,72699137,"dn"
90,#2022-03-29#,72699200,"up"
65,#2022-03-29#,72699247,"up"
```

Each record is assigned to a keyboard event and consists of four fields, separated by commas. In the first field, the virtual key code [21] of the key that participated in the keyboard event is recorded. IRecU records virtual key code values from 8 to 255, among which are all the keys on the keyboard, such as letters, numbers, punctuation marks, "Enter", "Alt", "Ctrl", "Shift", "Backspace", "Delete", etc., while those below 8 correspond to mouse actions. In the second field, the date on which the event took place is recorded. In the third field, the exact time when the event took place is recorded, expressed in the number of ms that have passed since the beginning of the day (12:00 a.m.). Finally, in the fourth field, the type of event is recorded, with "dn" corresponding to key press and "up" corresponding to key release.

According to them, from the above raw data recording example, the following can be concluded: (1) the keys with the codes "32", "90", and "65" were used, in order, which on



the English keyboard are assigned to “Space”, “Z”, and “A”, (2) the “65” key was pressed before the previous key, “90”, was released.

From the data recorded by IRecU, it can be known exactly what the volunteer typed and what day and time they typed it. It is understood that sensitive information of the volunteer is likely to be contained and therefore, as was also stipulated in the signed consent form, these data cannot be shared.

### 3.4. The Final Dataset

As is known, keystroke dynamics come with a large number of features, each of which contain a small amount of information. The most frequently used features in keystroke dynamics studies are keystroke durations and digram latencies [22].

Based on this observation and due to the fact that it is not permissible to share the raw data, according to the signed consent form, it was decided to extract the keystroke durations and down–down digram latencies from each logfile and make them available as a keystroke dynamics dataset. This dataset is named IKDD (IRecU’s Keystroke Dynamics Dataset) [23] and consists of several files, each of which were derived from a raw data dataset logfile. Each IKDD file includes the demographics of the volunteer recorded and a set of records, each of which maps to a keystroke dynamics feature and lists the values of that feature and that volunteer, in that particular logfile. Such a record has the following form:

$$x-y, \text{value1}, \text{value2}, \text{value3}, \dots$$

where  $x$  and  $y$  are the virtual key codes of the keys participating in the feature, and where  $\text{value1}$ ,  $\text{value2}$ ,  $\text{value3}$ , etc., are the values recorded for this feature. When  $y$  has the value 0, then the feature is keystroke duration, while when it has any other value, then the feature is down–down digram latency.

Some rules were followed for the extraction of the features. For example, regarding keystroke durations, values above 500 ms were rejected, based on the Windows key repeat rate preset. Also, with regard to digram latencies, values above 3000 ms were rejected, based on the fact that a time period greater than 3 s is considered by several studies as a typing pause [24].

An example of some records in an IKDD file is as follows:

```
48-0,62,65,74,64,60,45
49-0,95,91,82,108
50-0,98,88,87,103,104,59,87,65,60,48,83
69-82,272,316,671,391,96,928,550,74
69-83,125,193,170,142,235,168,310
69-84,180,604,362,409,171,147,190,158
```

The first field of each record indicates the feature. For example, the value “50–0” indicates the keystroke duration of the “2” key, while the value “69–84” indicates the digram latency of the “E–T” digram. All other fields are the values of the specific feature in the specific logfile. For example, the key “2” was used 11 times in this particular logfile. The first time this key was used, the keystroke duration recorded was 98 ms, the second time it was 88 ms, the third time it was 87 ms, and so on.

From the format of the IKDD files, it is understood that no sensitive or personal information of the volunteers can be revealed, and this is because it is not known in which order the keys and digrams were used, with the consequence that it is not possible to reconstruct the text, passwords, and credit card numbers.

## 4. Examples of Using IKDD

IKDD can be used in a variety of ways. Researchers can choose which keystroke durations and which digram latencies to use. They can also choose how the values of the features will be calculated. For example, the mean of the values of each record, or the

median, can be calculated as the value of the feature. Also, it can be chosen to include or not include outliers in feature value calculations, etc.

Next, some examples of the use of IKDD are listed, with the aim of demonstrating its utility. No further criteria are set, regarding, for example, the features that will be utilized, the classifiers that will be used, etc. The aim of the examples listed is not to find the machine learning model that leads to the best results, but to present the possibilities that IKDD offers to researchers. For this reason, some basic and well-known machine learning models are tested rather than more advanced ones. The process of the experiments performed is shown in Figure 2.

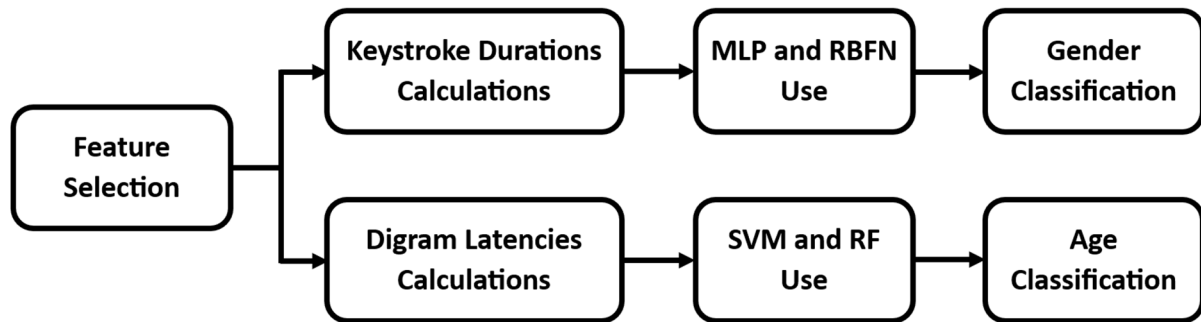


Figure 2. User classification flowchart.

#### 4.1. Recognizing the User's Gender

An example of using IKDD is to create a system to recognize a user's gender. As can be seen from Table 3, in the dataset, there are 279 logfiles from female users and 254 logfiles from male users. The values of the keystroke durations that are found in the IKDD files were used as data. For each key, the mean value of the keystroke durations was calculated, while if in a file a key was used less than five times, then it is considered that the sample is not representative and the value of the corresponding feature is considered unknown.

Two well-known neural networks, multi-layer perceptron (MLP) [25] and a radial basis function network (RBFN) [26], were used for classification. For each experiment that was performed, the 10-fold cross-validation method [27] was used, in which the dataset is divided into 10 equal-sized parts, of which 9 are used as a training test and 1 as a testing test, and this is repeated in a round robin mode.

Table 4 presents the results of the experiments performed to recognize the gender of the users. For each machine learning model, a number of experiments were performed to find the classifier settings that lead to the best results. By "best results", it meant the highest accuracy [28] and the shortest time for training the model (time to build model, TBM). In fact, the values of the F-measure (F1) and the area under the ROC curve (AUC) are also listed, which are alternative measurements of the performance of a system that are considered more reliable when the datasets are not balanced.

Table 4. Results of user's gender recognition.

Model	Acc. (%)	TBM (s)	F1	AUC
MLP	77.1	38.57	0.771	0.831
RBFN	81.2	1.14	0.812	0.851

As can be seen from Table 4, using only a few dozen keystroke dynamics features, without using any feature selection algorithm, and using a simple machine learning model, user gender recognition is achieved with a probability of more than 80%.

Among the two machine learning models, RBFN shows better performance. The confusion matrix corresponding to the highest accuracy is presented in Table 5.



**Table 5.** Confusion matrix of the highest accuracy in gender classification.

	Classified As	
	Male	Female
Male	207	47
Female	53	226

#### 4.2. Recognizing the User's Age Group

Another example of using IKDD is to create a system to recognize a user's age group. As shown in Table 3, in the dataset, there are 151 logfiles from users in the age group "18–25", 149 logfiles from users in the age group "26–35", 125 logfiles in the age group "36–45", and 108 logfiles in the "46+". The time digram latencies are chosen to be used as features. For each digram, the mean value of the digram latencies is calculated, while if in a file a digram was used less than three times, then it is considered that the sample is not representative and the value of the corresponding feature is considered unknown.

To achieve the goal, two well-known machine learning models were employed: the support vector machine (SVM) and the random forest (RF). Again, a number of experiments were performed to find the parameters of the classifiers leading to the best performing system, while the 10-fold cross-validation method was used again.

The results of the experiments are presented in Table 6.

**Table 6.** Results of user's age group recognition.

Model	Acc. (%)	TBM (s)	F1	AUC
SVM	70.5	0.96	0.705	0.838
RF	60.8	15.29	0.599	0.822

As shown in Table 6, a user's age group can be predicted with an accuracy of more than 70%, which is a significant improvement over random prediction. The percentage of random prediction is assumed to be 28%, which is the percentage of logfiles belonging to the class with the highest representation in the dataset.

The best performing machine learning model is SVM. The confusion matrix of the highest accuracy is presented in Table 7.

**Table 7.** Confusion matrix of the highest accuracy in age classification.

	Classified As			
	18–25	26–35	36–45	46+
18–25	116	17	12	6
26–35	20	97	21	11
36–45	12	20	82	11
46+	8	12	7	81

## 5. Discussion

One of the problems in the research field of keystroke dynamics is the lack of datasets from real user typing data. In fact, most of the already existing datasets are either fixed-text or contain a small amount of data, or contain data from a few volunteers. The keystroke dynamics dataset presented in this paper, named IKDD, was created from the daily computer typing of a large number of volunteers. Its use can help researchers to design systems for classifying users according to some of their inherent or acquired characteristics, in order to develop applications related to digital forensics, targeted advertising, ease of use of computing systems, the protection of unsuspecting users in cases of Internet fraud, etc.

IKDD has some advantages over other corresponding datasets. First, it contains data that best approximate actual user typing, because volunteers were allowed to type

whatever they wanted, whenever they wanted, in any computer application, for as long as it took, until a certain amount of data was completed. Second, the data come with five tags, thus allowing experiments to identify one or more user characteristics, including native language and educational level, tags that are rare in other datasets. Third, it contains a large amount of data, with each logfile having data from approximately 3500 keystrokes, thus adequately capturing the typing behavior of each volunteer.

IKDD contains data from the recording of 164 volunteers, from five different mother tongues, belonging to various educational levels, while gender, age group, and handedness are represented in the dataset with proportions that are also presented in the world population. For example, there are about as many females as males, while the ratio of left-handers to right-handers is 1 to 9. However, the dataset needs to be extended with the participation of more volunteers.

Therefore, the extension of this work will range in two levels. Firstly, other volunteers will be recruited whose daily keyboard usage will be recorded. Emphasis will be placed on capturing data that the IKDD lacks, specifically data from users of various native languages, particularly those that are widely spoken. Secondly, additional experiments will be conducted on user classification and ROC and AUC analysis will be performed, alongside the presentation of learning curves, with the aim, among other things, of examining possible under-fitting or over-fitting.

## 6. Conclusions

Keystroke dynamics is a biometric technology that can be used to authenticate users, to recognize certain inherent and acquired characteristics of users, and to recognize the physical and mental state of users. Keystroke dynamics studies require data obtained from recording the typing of volunteers. Due to the sensitive and/or personal information recorded, the datasets created in this way are not shared, with the result that there are very few keystroke dynamics datasets available on the Internet. This paper contributes an online-available keystroke dynamics dataset, named IKDD, which was created by recording the typing of 164 volunteers during their daily computer use. Furthermore, this paper describes the software developed and used to record keystrokes and create 533 logfiles. Also, the volunteers' participation consent form is described, in which the restrictions imposed on the researchers to protect the personal data of the volunteers are mentioned. The demographic characteristics of the dataset are given and finally some experiments are listed as examples of the use of IKDD.

IKDD contains data accompanied by five characteristics of the volunteers, namely gender, age group, handedness, mother tongue, and education level. Therefore, the limitations that follow are, firstly, that it can be used to classify users only in terms of these characteristics, while physical and mental state recognition research cannot be performed, and secondly, that researchers can only use keystrokes durations and digram latencies from the wide variety of keystroke dynamics features. Regarding the difficulties in creating the dataset, the most important ones were related to finding volunteers who would agree to have their typing recorded, risking the disclosure of sensitive and/or personal information.

The future goals of this work are the extension of the dataset, firstly, by recording more volunteers from more native languages, emphasizing the most widely spoken languages such as English, Chinese, Spanish, French, Arabic, etc., and secondly, by calculating and making available more keystroke dynamics features.

**Author Contributions:** Conceptualization, I.T.; methodology, I.T. and G.A.P.; software, I.T. and O.-D.A.; validation, I.T., O.-D.A., E.V. and G.A.P.; formal analysis, I.T. and O.-D.A.; investigation, I.T.; resources, I.T. and O.-D.A.; data curation, I.T., O.-D.A. and G.A.P.; writing—original draft preparation, I.T. and O.-D.A.; writing—review and editing, I.T., O.-D.A., E.V. and G.A.P.; visualization, O.-D.A.; supervision, I.T. and G.A.P.; project administration, G.A.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The IKDD is available at <https://github.com/MachineLearningVisionRG/IKDD>.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Ahmad, N.; Szymkowiak, A.; Campbell, P.A. Keystroke dynamics in the pre-touchscreen era. *Front. Hum. Neurosci.* **2013**, *7*, 835. [[CrossRef](#)] [[PubMed](#)]
2. Raul, N.; Shankarmani, R.; Joshi, P. A Comprehensive Review of Keystroke Dynamics-Based Authentication Mechanism. In *International Conference on Innovative Computing and Communications*; Khanna, A., Gupta, D., Bhattacharyya, S., Snasel, V., Platos, J., Hassanien, A.E., Eds.; Advances in Intelligent Systems and Computing; Springer: Singapore, 2020; Volume 1059, pp. 149–162. [[CrossRef](#)]
3. Buker, A.A.N.; Roffo, G.; Vinciarelli, A. Type Like a Man! Inferring Gender from Keystroke Dynamics in Live-Chats. *IEEE Intell. Syst.* **2019**, *34*, 53–59. [[CrossRef](#)]
4. Pentel, A. Predicting User Age by Keystroke Dynamics. In *Artificial Intelligence and Algorithms in Intelligent Systems*; Silhavy, R., Ed.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2019; Volume 764, pp. 336–343. [[CrossRef](#)]
5. Monaro, M.; Spolaor, R.; Li, Q.; Conti, M.; Gamberini, L.; Sartori, G. Type Me the Truth! In Proceedings of the ARES'17: International Conference on Availability, Reliability and Security, Calabria, Italy, 29 August–1 September 2017; p. 60.
6. Epp, C.; Lippold, M.; Mandryk, R.L. Identifying emotional states using keystroke dynamics. In Proceedings of the CHI '11: CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 715–724.
7. Marrone, S.; Sansone, C. Identifying Users' Emotional States through Keystroke Dynamics. In Proceedings of the 3rd International Conference on Deep Learning Theory and Applications, Lisbon, Portugal, 2–14 July 2022; pp. 207–214.
8. Kołakowska, A.; Landowska, A. Keystroke Dynamics Patterns While Writing Positive and Negative Opinions. *Sensors* **2021**, *21*, 5963. [[CrossRef](#)] [[PubMed](#)]
9. Crawford, H. Keystroke dynamics: Characteristics and opportunities. In Proceedings of the 2010 Eighth Annual International Conference on Privacy, Security and Trust (PST), Ottawa, ON, Canada, 17–19 August 2010; pp. 205–212.
10. Tsimperidis, I. User Classification Through Keystroke Dynamics, for Suspect Identification, Democritus University of Thrace, Xanthi, Greece. 2017. Available online: <https://www.didaktorika.gr/eadd/handle/10442/40524> (accessed on 20 May 2024).
11. Killourhy, K.S.; Maxion, R.A. Comparing anomaly-detection algorithms for keystroke dynamics. In Proceedings of the Networks (DSN), Lisbon, Portugal, 27–30 June 2023; pp. 125–134.
12. Loy, C.C.; Lai, W.K.; Lim, C.P. Keystroke Patterns Classification Using the ARTMAP-FD Neural Network. In Proceedings of the Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kaohsiung, Taiwan, 26–28 November 2007; pp. 61–64.
13. Dias, T.; Vitorino, J.; Maia, E.; Sousa, O.; Praça, I. KeyRecs: A keystroke dynamics and typing pattern recognition dataset. *Data Brief* **2023**, *50*, 109509. [[CrossRef](#)] [[PubMed](#)]
14. Risto, H.N.; Graven, O.H. Collection and Statistical Analysis of a Fixed-Text Keystroke Dynamics Authentication Data Set. In Proceedings of the 2023 7th Cyber Security in Networking Conference (CSNet), Montreal, QC, Canada, 16–18 October 2023; pp. 67–73.
15. Giot, R.; El-Abed, M.; Rosenberger, C. GREYC keystroke: A benchmark for keystroke dynamics biometric systems. In Proceedings of the 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS), Washington, DC, USA, 28–30 September 2009; pp. 1–6.
16. Sznur, S.; García, S. Advances in Keystroke Dynamics Techniques to Group Users Sessions. *Int. J. Inf. Secur. Sci.* **2015**, *4*, 26–38.
17. Maalej, A.; Kallel, I. Does Keystroke Dynamics tell us about Emotions? A Systematic Literature Review and Dataset Construction. In Proceedings of the 2020 16th International Conference on Intelligent Environments (IE), Madrid, Spain, 20–23 July 2020; pp. 60–67. [[CrossRef](#)]
18. Vural, E.; Huang, J.; Hou, D.; Schuckers, S. Shared research dataset to support development of keystroke authentication. In Proceedings of the 2014 IEEE International Joint Conference on Biometrics (IJCB), Clearwater, FL, USA, 29 September–2 October 2014; pp. 1–8.
19. El-Abed, M.; Dafer, M.; El Khayat, R. RHU Keystroke: A mobile-based benchmark for keystroke dynamics systems. In Proceedings of the 2014 International Carnahan Conference on Security Technology (ICCST), Rome, Italy, 13–16 October 2014; pp. 1–4.
20. Güntürkün, O.; Ströckens, F.; Ocklenburg, S. Brain Lateralization: A Comparative Perspective. *Physiol. Rev.* **2020**, *100*, 1019–1063. [[CrossRef](#)] [[PubMed](#)]
21. Zyrianov, V.; Peterson, C.S.; Guarnera, D.T.; Behler, J.; Weston, P.; Sharif, B.; Maletic, J.I. Deja Vu: Semantics-aware recording and replay of high-speed eye tracking and interaction data to support cognitive studies of software engineering tasks—Methodology and analyses. *Empir. Softw. Eng.* **2022**, *27*, 1–39. [[CrossRef](#)] [[PubMed](#)]
22. Tsimperidis, I.; Arampatzis, A. The Keyboard Knows About You: Revealing User Characteristics via Keystroke Dynamics. *Int. J. Technoethics* **2020**, *11*, 34–51. [[CrossRef](#)]

23. IKDD (IRecU's Keystroke Dynamics Dataset). Available online: <https://github.com/MachineLearningVisionRG/IKDD> (accessed on 21 July 2024).
24. Alves, R.A.; Castro, S.L.; Olive, T. Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *Int. J. Psychol.* **2008**, *43*, 969–979. [[CrossRef](#)] [[PubMed](#)]
25. Alnuaim, A.A.; Zakariah, M.; Shukla, P.K.; Alhadlaq, A.; Hatamleh, W.A.; Tarazi, H.; Sureshbabu, R.; Ratna, R. Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier. *J. Health Eng.* **2022**, *2022*, 6005446. [[CrossRef](#)] [[PubMed](#)]
26. Zhang, R.; Li, Y.; Gui, Y.; Zhou, J. Prediction of blasting induced air-overpressure using a radial basis function network with an additional hidden layer. *Appl. Soft Comput.* **2022**, *127*, 109343. [[CrossRef](#)]
27. Wieczorek, J.; Guerin, C.; McMahon, T. K-fold cross-validation for complex sample surveys. *Stat* **2022**, *11*, e454. [[CrossRef](#)]
28. Hao, S.; Xu, H.; Ji, H.; Wang, Z.; Zhao, L.; Ji, Z.; Ganchev, I. G2-ResNeXt: A Novel Model for ECG Signal Classification. *IEEE Access* **2023**, *11*, 34808–34820. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.