



# Chapter 1

## User Profiling Using Keystroke Dynamics and Rotation Forest

**Ioannis Tsimperidis**

 <https://orcid.org/0000-0003-0682-1750>  
*Democritus University of Thrace, Greece*

**Avi Arampatzis**

 <https://orcid.org/0000-0003-2415-4592>  
*Democritus University of Thrace, Greece*

### **ABSTRACT**

*The anonymity that users can maintain when connecting to the internet, in addition to the positive effects, such as being able to express their views and ideas freely without fear of retaliation, also carries some risks, such as the fact that it is a significant advantage for malicious users. In order to remove the complete anonymity of internet users, so as to protect unsuspecting users, this work attempts to identify some of their characteristics, namely gender, age, and handedness, using data coming from typing. For this purpose, the rotation forest is used as a classifier, and keystroke dynamics features are selected based on the chi-square feature selection procedure. The final results show that user profiling can be achieved with an accuracy of 88.9% in gender prediction, 86.3% in age prediction, and 94.3% in handedness prediction.*

### **INTRODUCTION**

People work, communicate, trade goods and services, are entertained and educated, and much more, in a very different way than a few years ago. Telecommunication and teleconferencing applications, various eShops, online games, courses of any

DOI: 10.4018/978-1-7998-9430-8.ch001

kind, and many more, have made their appearance serving the needs of individuals, companies, and organizations. The cause of all these rapid changes is the evolution and dissemination of the Internet and the services it offers. Today, a user has the ability to connect with other users from anywhere in the world through video calling or instant messaging applications, or through social networks. Also, every user has the opportunity to purchase products or services from the global market, with the same ease that he/she would do in his/her neighborhood, or even easier. It is also possible to find work for or with companies and individuals that may be located thousands of kilometers away.

Many opportunities for personal, national, and global growth and development are offered, but at the same time there are many risks, such as financial frauds, seduction of minors, hacking, anonymous threats, etc. (Degtereva et al., 2020). One of the most important reasons for the existence of these risks is the partial or complete anonymity that a user can maintain when connecting to the Internet. This anonymity, on the one hand, often proves useful as it helps the user to express and freely be creative, but on the other hand may alter his/her behavior by turning him/her into a rude, aggressive, and disrespectful person (Krysowski & Tremewan, 2020). In addition, anonymity or concealment of true identity is one of the major advantages of malicious users in their plans to deceive unsuspecting users and/or carry out cyber-attacks.

Also noteworthy is that the way in which users interact on the Internet is shaped by the fact that although a variety of communication methods are offered, such as voice calls, video calls, file sharing, etc., text is still the dominant form of communication (Nitzburg & Farber, 2019) among users. A variety of instant messaging applications are available and many companies invest significant amounts of money in their development. If we additionally consider the email service, the comments made by users on various social media, and searches carried out in search engines, each of which is primarily in text, a backdrop is formed in which text, or rather text typing, plays a prominent role on the World Wide Web, in user communication, and in computer operations in general.

Keystroke dynamics are a biometric trait from which information can be extracted by exploiting data that comes from the way a user types on a real or virtual keyboard. Studies in keystroke dynamics have been conducted for about fifty years and their object is mainly user authentication (Raul et al., 2020) in order to replace or enhance the authentication method using passwords. Keystroke dynamics were also used to classify users according to an inherent or acquired characteristic, such as gender or age, as well as to assess users' physical and mental condition, such as whether they were exhausted (Ulinskasa et al., 2018), if they suffer from depression (Mastoras et al., 2019), or if they suffer from a neurological disease (Lam et al., 2020). Many

years of experimentation with keystroke dynamics resulted in the development of systems with very good performance in user authentication and user classification.

The features used in keystroke dynamics relate to how the users type, not what they type, and can be divided into temporal and non-temporal. The most commonly used temporal features are the keystroke durations which are the time intervals between the push of a button and its release, and the digram latencies which are the time intervals between the uses of two consecutive keys. Digram latency can be expressed in four different ways, firstly, the time elapsed between the pressing of the first key and the pressing of the second key (i.e. the down-down digram latency or DDDL), secondly, the time elapsed between the pressing of the first key and the releasing of the second key (i.e. the down-up digram latency or DUDL), thirdly, the time elapsed between the releasing of the first and the pressing of the second key (i.e. the up-down digram latency or UDDL), and fourthly, the time elapsed between the releasing of the first and the releasing of the second key (i.e. the up-up digram latency or UUDL). Other temporal keystroke dynamics features are trigram latencies, which are similarly defined, tetragram latencies, and generally n-gram latencies, the number and duration of typing pauses, typing speed, etc. Non-temporal features include the percentages of usage of each of the duplicate keys, such as “Shift”, “Ctrl”, and the number keys, the mode of correction of typing errors (i.e., backspace vs. delete), the application in which typing is performed, and other typing features.

This work uses keystroke dynamics to find some characteristics of completely unknown Internet users, in order to remove complete anonymity and solve some of the problems mentioned. To achieve this, a machine learning model is used which combines simplicity of operation with efficiency, namely the rotation forest. The next section of the chapter provides a review of the literature related to the topic of user classification using keystroke dynamics. Then an analysis is made of the stages of the methodology followed and the results of the experiments conducted are presented. Following are suggestions for exploiting the findings of this chapter and references to possible extensions of this research. Finally, the conclusion of the chapter is presented.

## **KEYSTROKE ANALYSIS**

The reasons for trying to identify certain characteristics of a computer user vary. For example, when a cybercrime is committed and the culprit is sought, it would be a valuable help if some of his/her characteristics were known, such as gender, age, handedness, mother tongue, educational level, etc., in order to reduce the number of suspects. In another application, targeted advertising would benefit, since on the one

hand the investment of the companies would have better results and on the other hand the users would not be overwhelmed with many and indifferent advertising messages, but with much less and more targeted ones. Also, knowing some characteristics of the user using a computer enables the user-computer interaction to become much more successful. That is, it would be possible to provide advice and suggestions to the users to visit certain websites, use certain services, and participate in certain groups that are more suitable for them. In addition, by revealing the characteristics of the users, it would be possible to warn unsuspecting users about the possibility of falling victim to some deception. These reasons, and possibly others, are the motivation for many works in the field of keystroke dynamics aimed at identifying certain characteristics of computer and Internet users.

In one such study, Fairhurst and Da Costa-Abreu (2011) focused on the use of social networks by young people and the existence of risks of hiding the real characteristics of users. They used an existing dataset with data from 98 male and 35 female users. They used three simple classifiers, namely k-nearest neighbors, C4.5 decision tree, and naive Bayes, as well as three classifier combination techniques. The best results came from the dynamic classifier selection based on local accuracy class (DCS-LA) (Woods et al., 1997), with an error rate of 3% in gender prediction.

Antal and Nemes (2016) attempted to identify the user gender of a mobile device from data from touchscreen swipes and keystroke dynamics. Thus, they used two datasets, one created by recording touchscreen swipes while answering a questionnaire consisting of 58 questions and one created by recording the typing of a specific password by 42 users, 24 males and 18 females. In terms of keystroke dynamics the features that were extracted were the keystroke durations, the down-down digram latencies, the pressure exerted on the virtual keys, and the surface covered when using the keys. Random forest was used for the classification and the results showed an accuracy of 93.5% in the identification of the user's gender.

Lee et al. (2018) also dealt with mobile devices in their research and aimed to solve the problem of authentication of smartphone users using PIN or pattern drawing, due to the fact that it is very vulnerable to the shoulder surfing attack. They collected data from typing on smartphones and as features, among others, used keystroke durations and all versions of digram latencies. Researchers using distance algorithms have been able to identify an impostor with an equal error rate (EER) of 8%, which is an indicator of system performance. EER is a point where false acceptance rate and false rejection rate intersects, and the lower it is, the more accurate the system. But importantly, they also found that it is easier to identify an impostor when the legal user is of the opposite gender, thus proving that it is possible to separate users according to their gender depending on the way they type, offering another suggestion for implementing gender classification.

Udandarao et al. (2020) examined the effect of various characteristics of the users on the way they type, such as their computer experience, their gender, their height, etc. They used an existing dataset created by recording 117 volunteers, who typed two specific sentences and answered a series of questions. Regarding the gender demographics of the volunteers, 72 were males and 45 were females. The features they used were keystroke durations, all types of digram latencies, as well as features related to whole words. For the gender classification, six machine learning models and four deep learning models were tested, of which the convolutional neural network (CNN) achieved the highest accuracy of 93%.

Identifying the gender of the person we are talking to is a simple process during a face-to-face conversation. Facial characteristics, expressions, and differences imposed by some cultures (such as hairstyle and clothing), are clues for making such an identification. But all these are absent when chatting on the Internet and this is the reason why Buker and Vinciarelli (2021) conducted their research to reveal the gender of the user who communicates via chat applications. They collected data from the discussion of 60 people, in pairs, through a chat application, of which 35 were females and 25 were males. The features extracted were the density of “!”, density of “?”, density of non-alphabetic characters, typing speed, backspace time, etc. For classification they used a random forest reaching an accuracy of 98.8% and showing what the most important features are for separating users according to their gender, i.e. typing speed, backspace time, and density of “backspaces”, among others.

Data similar to those derived from keystrokes were addressed by Van Balet et al. (2016). The reason for their research was that gender is hidden in online conversations often for malicious purposes. Given the differentiation of gestures between males and females, the possibility of separating users according to their gender depending on mouse movements was examined. Data were collected from 94 users (49 women and 45 men) with the two groups having similar statistical characteristics in terms of age and computer experience. Features were extracted from the data such as the time that the left click remains pressed, the maximum speed observed during the movement of the mouse, the total distance traveled by the mouse during an action, etc. The user gender was predicted using logistic regression and the results showed an accuracy approaching 76% in an independent test set and once the outliers have been removed.

Gender is the user characteristic sought in most keystroke dynamics research, mainly because it is a characteristic that is quite distinct compared to others, such as age and educational level, as well as because it seems to be of the greatest commercial interest. Idrus et al. (2014) in their work dealt with other characteristics besides gender and attempted to do user profiling from data coming from keystrokes. For this reason they used two datasets, one created by recording the typing of five short phrases and one by recording free text typing. As a classifier they used a support vector machine

(SVM) with a radial basis function kernel. They performed experiments to find the gender, the age group (<30 and <sup>3</sup>30), and the handedness of users. In each group of experiments the datasets were balanced by removing excess instances, so that for example the number of males is equal to the number of females, etc. The results showed accuracy up to 86% in gender classification, up to 78% in age classification, and up to 88% in handedness classification.

Beyond gender, the second most frequently sought characteristic is age. In the studies that involve classification, age groups are defined, and the aim is to find the group to which a user belongs. The segregation of groups in each separate research has been usually done based on the limits set by legislation, such as the age that separates minors from adults, based on the available dataset so that classes with the same number of instances emerge, or based on other criteria, which could also be arbitrary. Thus, Tsimperidis et al. (2021) arbitrarily defined four classes and found the age group that a user belongs to, utilizing data from the typing patterns. A dataset from the typing recording of 118 volunteers was used and keystroke durations and down-down diagram latencies were used as features. Of the five classifiers tested, radial basis function network (RBFN) was the most successful with 90% accuracy.

The dangers of the Internet for children led Uzun et al. (2015) to check how successfully typing data can be used to distinguish children from adults. For the needs of their research, they recorded the typing of users who belonged to two age groups, 10-14 and 18-49 years old. The recording was made on a specific computer with an application created by the researchers in which they invited the volunteers to answer some questions. For the separation between children and adults they used a number of classifiers, of which the linear SVM proved to be the most successful with EER 8.8%.

In their work, Hossain and Haberfeld (2020) attempted to separate children from adult users, again with the aim of protecting minors from the dangers of the Internet. They focused on mobile devices and created an application for recording users, in which volunteers were asked to press six keys, in a specific order, several times. They divided users into three age groups, children (5-12 years old), adolescents (13-17 years old), and adults (18 years old and older). The features that were taken advantage of were keystroke durations, the surface occupied by the finger, and the pressure exerted on each virtual key. For the classification they used linear models, nearest neighbors, and SVMs. The results showed a successful identification of the user's age group with a percentage of about 73% on smartphones and 82% on tablets.

In another work, Vesel et al. (2020) was trying to find out if there are any obvious differences in the way users with mood disorders type. For this purpose, they used keystroke dynamics data in order to diagnose depression, bipolar disorder, anxiety, attention deficit hyperactivity disorder, post-traumatic stress disorder, etc. As features they used inter-key delay (IKD) which are the DDDLs between key types (not between

### ***User Profiling Using Keystroke Dynamics and Rotation Forest***

each individual key), typing speed, and pauses during typing. An important finding of their research is the significant differentiation in IKDs between the age groups of individuals close to 20 years, close to 45 years, and close to 70 years, which makes it possible to separate users according to their age depending on the way they type.

User handedness is a characteristic which is rarely explored in research, mainly due to the fact that the datasets that are created are extremely unbalanced and the classification procedure is very difficult. In a study, Roy et al. (2018) attempt to reveal the handedness of a smartphone user, among other characteristics. To create a keystroke dynamic dataset, they developed a web-based application and recorded 92 users typing a particular word seven times. Keystroke durations and all types of digram latencies were used as features. After removing the outliers from their data, they proceeded to classification using the SVM, naive Bayes, random forest, and multinomial nominal log linear model. The best results came from random forest with 81.5% accuracy.

The handedness of an unknown user, among other characteristics, is sought in the work of Tsimperidis et al. (2021). Researchers recorded typing by a number of volunteers during the daily usage of their computers. From the data they collected they extracted 230 keystroke durations and digram latencies, and by testing five different machine learning models they were able to identify the dominant hand of an unknown user with 97% accuracy. While in another study, Earl et al. (2021) tried to show that the combined use of keystroke and mouse dynamics features can bring better results in recognizing some user characteristics. To collect keystroke data 240 volunteers copied a piece of text and answered a question. Digram latencies and the error rates were extracted from the recorded data as features. They followed a feature selection process and tried some combinations of features to achieve the best results. Decision trees, random forest, Gaussian naive-Bayes, SVM, and K-nearest neighbors were used for classification. Finally, experiments showed that a user's handedness can be predicted with 73.5% accuracy taking advantage of keystroke dynamics features.

In addition to the gender, age, and handedness that are sought in this chapter, in the literature there are also studies that aim to find other user characteristics, such as ethnicity, educational level, etc. The bottom line, however, is that more and more efforts are being made in this direction, with new techniques being tested, and there are now quite reliable systems for finding certain characteristics of computer users by exploiting data derived from typing.

## **METHODOLOGY AND RESULTS**

The methodology followed in the present study consists of three steps. Firstly, the keystroke dynamics data collection. Secondly, the extraction of features from the data and the selection of the most appropriate ones for user classification according to the gender, age, and handedness. Thirdly, the use of a machine learning model and finding its appropriate parameters for effective user classification.

### **Data Acquisition**

An appropriate dataset in keystroke dynamics studies is crucial for performing experiments and drawing correct conclusions. The dataset should be accompanied by the appropriate demographics and contain the required data. In some keystroke dynamics research ready-made datasets were used (Giot et al., 2015) while in others new ones had to be created. Creating a keystroke dynamics dataset can be done by recording the typing of users who have been asked to copy a specific piece of text, a task usually performed in a closed environment, or by recording the typing of users who type at will, something that is done either by answering specific questions and performing specific tasks, or by using the computer without any restrictions and instructions. The former way to create a keystroke dynamics dataset is called fixed-text and the latter free-text.

In the data acquisition task of this work, the mode that approaches the normal operation of the computer as close as possible was selected. Specifically, a keylogger was installed on the volunteers' computer which has the ability to record typing actions from any application in a Windows environment. For security and privacy reasons, the volunteers were given the opportunity to enable and disable the keylogger whenever they wished, to monitor the recorded data but without being able to modify it, to leave the process at any time, and to decide whether to deliver or not the log files. In addition, a consent form was signed in which the researchers pledged not to share the data in any way and to use it only for the purposes of the present study. This ensures that personal and/or sensitive data, such as passwords and personal messages are not leaked to other people.

In a period that lasted a little over 18 months, 118 volunteers were recorded and handed over log files. Each of the volunteers submitted 3-4 log files resulting in the creation of a dataset of 387 log files, each of which contains data of approximately 3,500 keystrokes and metadata with the characteristics of the volunteers, among which was the gender, the age, and the handedness of users. In each log file, each record corresponds to an action on the keyboard and consists of four fields, separated by a comma. The first field lists the key on which the action was performed, in the form of a virtual key code, which is a standard encoding by which each key and each mouse



action is assigned a number from 1 to 255. With the first 7 codes corresponding to mouse actions the recording concerned codes 8 to 255. The second field records the date on which the action took place. The third field records the exact time that the action took place in the form of an integer that indicates the number of milliseconds that have elapsed since the beginning of the day, i.e., at 12 midnight. Finally, the fourth field lists the type of action, which can be key-press or key-release.

## **Feature Extraction and Feature Selection**

From the data recorded in the dataset it is possible to extract most of the features used in keystroke dynamics studies. For example, subtracting the value in the third field of a record for one keypress from the corresponding value of the next record for the key-release of the same key results in the keystroke duration. Also, subtracting the values of the third field that have two consecutive records for key presses results in DDDL. Moreover, counting the number of records that first field has the value 160, and those that have the value 161, results in the number of times the left and right “Shift” were used, respectively, and therefore the percentage of use of each of these. In similar ways it is possible to extract many other keystroke dynamics features.

The number of available features is in the order of millions and therefore a choice must be made as to which of them will be used. As such, in the present work, the most widely used features in keystroke dynamics studies, i.e., keystroke durations and down-down digram latencies, were selected. In a log file, each key and each digram has been recorded many times, resulting in many different measurements for the same feature. Finally, the value of the feature is the average of these many measurements. In fact, for reasons of reliable calculation of feature values, when the use of a key in a log file has been recorded less than five times it is not taken into account. Similarly, a digram latency is not considered if the corresponding digram has been recorded less than three times.

Approximately 65,000 features were extracted with this process, which is a very large number and the use of all of them will lead to time consuming systems. For this reason, a feature selection procedure was followed in order to find those features that can best distinguish the users according to gender, age, and handedness.

The Chi-Square feature selection was followed as such procedure. In feature selection, Chi-Square calculates the correlation between a class and a feature. When the resulting value of the Chi-square is small it means that it will be difficult to distinguish the classes only using the feature as class differentiator, and therefore may be rejected. On the contrary, when the value is high then this feature is characterized as capable of separating classes. The problem is that the Chi-Square feature selection procedure can be applied when classes and features are categorical, but the features used in this classification are measured in milliseconds, i.e., they are numerical.

However, if numerical features are suitably discretized, they can also be used in the procedure.

The Chi-Square value for each feature  $f$ , which has discretized in  $v$  values, in a classification problem with  $C$  classes, is given by the formula:

$$\chi^2(f) = \sum_{i=1}^v \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

$O_{ij}$  is the number of times where feature  $f$  is observed to have the  $i$ -th value in the  $j$ -th class.  $E_{ij}$  is the number of times where feature  $f$  is expected to have the  $i$ -th value in the  $j$ -th class. If this procedure, which is also described in the work of Rachburee and Punlumjeak (2015), is followed for each feature in each of the three classification problems, three lists of features are generated, with features ranked by their usefulness (as measured by their Chi-Square values) in separating users according to the characteristics studied.

In Table 1, the first 15 features are ranked with the highest Chi-Square for gender, age, and handedness classification problems, where each of them is represented by the virtual key code(s) of the keys that compose it. So, one number indicates keystroke duration, and two numbers indicate down-down digram latency.

Some observations that can be made from Table 1 are: a) keystroke durations seem to play a more important role than digram latencies in age classification, while digram latencies are more significant in gender classification; b) the “A”, “M”, “N”, and “O” keys, along with the digrams in which they participate, show high correlation with gender; c) the keys “A” and “T”, along with the digrams in which they participate, are placed quite high on the list of the most important features in case of handedness classification; d) in case of age classification, Chi-Square values are much higher than the other two classification problems, which means that a feature that is in the two, or all three, lists, in the same ranking position, as for example with the “D” keystroke duration, is more capable of separating users by age than others characteristics. It should be noted that as far as observation (b) is concerned, the keys mentioned are at the left end and at the right end of the character range on the keyboard. This finding needs to be further studied to find out if there is a correlation between the location of the keys on the keyboard and the way they are used by users with different characteristics.

## User Profiling Using Keystroke Dynamics and Rotation Forest

Table 1. Keystroke dynamics features with the highest Chi-Square in gender, age, and handedness classification

#	Gender			Age			Handedness		
	Feat.	Keys	$\chi^2$	Feat.	Keys	$\chi^2$	Feat.	Keys	$\chi^2$
1	80-65	P-A	28.4685	65-32	A-(space)	79.2127	79	O	51.0753
2	77-65	M-A	25.8396	69	E	71.9764	84-65	T-A	39.4705
3	73-78	I-N	24.0143	79	O	50.5387	82-65	R-A	35.2808
4	78-65	N-A	23.8809	65	A	43.8988	71	G	28.8362
5	68	D	22.4733	68	D	41.9577	65	A	28.7747
6	77-79	M-O	21.7047	83	S	40.6108	65-84	A-T	25.2559
7	75-65	K-A	21.4770	32	(space)	40.5972	186	::	23.4427
8	78-79	N-O	20.5897	87	W	40.4034	83-84	S-T	21.3273
9	76-69	L-E	20.1416	39	(right-arrow)	39.2502	69	E	21.1344
10	79-77	O-M	19.4597	89	Y	36.7346	76-69	L-E	20.3488
11	65	A	19.2735	86	V	36.0397	66	B	19.0795
12	69-73	E-I	18.8601	70	F	35.3603	65-32	A-(space)	17.7110
13	79-78	O-N	18.7485	88	X	33.7660	82	R	16.0820
14	65-83	A-S	18.5307	73-32	I-(space)	33.6719	186-89	::-Y	15.9355
15	87	W	18.4606	78	N	30.4884	76-73	L-I	14.6795

## Rotation Forest

As mentioned in the section “Keystroke Analysis”, many classifiers have been used to classify users using keystroke dynamics features. Among them are SVM, random forest, naïve Bayes, RBFN, k-nearest neighbors, C4.5 decision tree, and many others. One classifier that has not been used so far in keystroke dynamics user classification studies, to our knowledge, is the rotation forest. It is a classifier ensemble where each base classifier uses a different training set, and all of them can be trained in parallel.

For each classifier in the ensemble the available feature set, for each of the classification problems, is divided into a number of subsets. These subsets may be disjoint or intersecting, but to achieve greater diversity in the training sets of base classifiers the disjoint subsets are preferred. For each of the feature subsets, half the classes of the problem are randomly selected and only the instances labeled with these classes are retained from the original dataset. Thus, a number of different sub-datasets are created. For each of these sub-datasets a percentage of the remaining instances is randomly removed. Then, principal component analysis (PCA) is performed on

the features and the instances of each sub-dataset to calculate the coefficients of principal components and to form a sparse matrix. The columns of this matrix are rearranged to correspond to the original features. Finally, the training set for a base classifier is calculated by multiplying this “rotation” matrix with the initial dataset. When this process is completed, the training set for each base classifier is created. This algorithm is described in more detail in the work of Rodriguez et al. (2006).

Therefore, the classifier parameters are the number of base classifiers, the number of features that will form a subset (which can be set between two values), and the percentage of instances that are removed from the dataset. It is noted that the C4.5 decision tree is chosen as the base classifier, on the one hand because of its simplicity, and on the other hand because it is sensitive to rotation of the features.

## **Experiments and Results**

The keystroke dynamics feature selection procedure showed 514, 690, and 246 features with a non-zero Chi-Square value for gender, age, and handedness classification problems, respectively. In the experimental procedure that was followed, all these features were used, and a different number of base classifiers were tested. Specifically, experiments were conducted for 10, 20, 30, 40, and 50 base classifiers, and for each different number the best performance of the rotation forest was sought, as measured by the accuracy, the training time (time to build model, TBM), the F-score (F1), and the area under the ROC curve (AUC).

The F1 score is used, as a combined measurement of precision and recall, because accuracy alone cannot fully give the picture of the overall performance of a model when classes are imbalanced, and because the F1 score is a measurement of how balanced the prediction across classes is. For example, assume two cases of a system for a handedness classification problem, where, as expected (Papadatou-Pastou et al., 2020), the ratio of left versus right handers is 1:10. In the first case, the system predicts all users as right-handed. The accuracy is 90%, but it is obvious that the system is not working properly. In the second case, the system correctly predicts the dominant hand of users 9 out of 10 instances, for all classes. The accuracy is again 90%, but this system is more reliable. This greater reliability is reflected in the F-score, where in the latter case is higher.

AUC, which is a common tool for evaluating predictions, e.g., Cook and Ramadas (2020), is also used to form a more complete picture of classifier performance. The receiver operating characteristic (ROC) curve is a plot that presents the recall as a function of probability of false alarm, which is equal to  $1 - \text{precision}$ . The ROC curve is limited to the interval  $[0, 1]$  in both dimensions, thus AUC, which is an area enclosed between the curve and the false positive rate axis, varies between 0 and 1.

The well-known 10-folds cross-validation was used in the experiments, i.e., the dataset is randomly divided into 10 disjoint parts with approximately equal size and every part is in turn used to test the model induced from the other 9 parts, e.g., Wong and Yang (2017). In this study, where there are 387 log files, each part in which the dataset was divided consists of 38 or 39 files. With the volunteers having delivered 3-4 log files it was easy to include all files of each individual in one of the 10 parts, so that to avoid overfitting in case that one log file from a person could end up in the training set while another one ends up in the testing set.

## Gender Classification

In the gender classification problem two classes were defined, “male” and “female”, and out of the 118 volunteers who participated in the typing process, 61 were male (51.7% of all volunteers) who submitted 203 log files (52.4% of all log files) and 57 were female (48.3% of all volunteers) who submitted 184 log files (47.6% of all log files). That is, the dataset is gender balanced and reflects global demographics, since men and women are roughly equal in number. Table 2 shows the best performance of the rotation forest for different numbers of base classifiers.

The best performance shown in Table 2 was achieved for the case of 10 base classifiers having subsets between 3 and 12 features and removing 50% of instances, for the case of 20 base classifiers having subsets between 1 and 10 features and removing 25% of instances, for 30 base classifiers the rotation forest parameters were the creation of subsets between 5 and 10 features and the removal of 10% of the instances, in the case of 40 base classifiers the values of respective parameters were 9, 10, and 50%, and finally, for the case of 50 C4.5 decision trees, which is the base classifier, the corresponding values of the rotation forest parameters were 3, 12, and 25%.

*Table 2. Performance of the rotation forest in the gender classification problem for different numbers of base classifiers*

Base Classifiers	Acc.	TBM (secs)	F1	AUC
10	85.0%	3.27	0.850	0.916
20	88.1%	9.33	0.881	0.936
30	87.9%	14.69	0.879	0.939
40	88.4%	14.19	0.884	0.953
50	88.9%	23.09	0.889	0.950

An obvious conclusion drawn from Table 2 is that performance seems to increase as more base classifiers are used, at a cost of increasing TBM.

## Age Classification

Four age classes were defined in the age classification problem, “18-25”, “26-35”, “36-45”, and “46+” years old users. Of the 118 volunteers, 31 belonged to the age group “18-25” (26.2% of all volunteers) who submitted 96 log files (24.8% of all log files), 37 belonged to the age group “26-35” (31.4%) who submitted 129 log files (33.3%), 37 belonged to the age group “36-45” (31.4%) who submitted 117 log files (30.2%), and 13 belonged to the age group “46+” (11.0%) who submitted 45 log files (11.7%). The dataset is balanced in terms of the first three classes, while the fourth class is less represented, although the number of instances is considered sufficient as it is less than three times smaller than that of the other classes. Table 3 shows the best performance of the tested classifier for 10, 20, 30, 40, and 50 C4.5 decision trees.

In Table 3 the best performance of the rotation forest with 10 base classifiers was achieved creating subsets having features between 9 and 10 and removing 50% of instances, while with 20 decision trees was achieved with subsets of 10 to 15 features and removing the 75% of instances, with 30 decision trees with subsets of 9 to 10 features and removing the 75% of instances, with 40 trees with 10 to 15 features in subsets and removing 90% of instances, and finally, in the case of 50 base classifiers the best performance achieved having subsets of 10 to 15 features and removing the 50% of instances.

The conclusion drawn from Table 3 for age classification, similar to that of gender classification in Table 2, is that effectiveness seems to increase as more base classifiers are used, at a cost of increasing TBM.

*Table 3. Performance of rotation forest in the age classification problem for different numbers of base classifiers*

Base Classifiers	Acc.	TBM (secs)	F1	AUC
10	80.1%	6.14	0.799	0.927
20	83.2%	8.69	0.830	0.953
30	83.5%	12.88	0.833	0.951
40	85.0%	13.60	0.848	0.951
50	86.3%	29.92	0.862	0.963

## Handedness Classification

In most keystroke dynamics studies dealing with handedness, two classes were defined, “right-handed” and “left-handed”, as shown in the section “Background”. But in this research the class “ambidextrous”, in which included users who said they use both right and left hand with the same skill, is added. Of the 118 volunteers who participated in the process, 105 were “right-handed” (89.0% of all users) who submitted 343 log files (88.6% of all log files), 10 were “left-handed” (8.5%) who submitted 35 log files (9.0%), and 3 were “ambidextrous” (2.5%) who submitted 9 log files (2.4%). The dataset is as unbalanced as would be expected according to global demographics. Table 4 presents the best performance of rotation forest for different numbers of base classifiers in predicting the dominant hand of users.

The values of rotation forest parameters, and specifically the minimum number of features in each subset, the maximum number of features, and the percentage of instances removed, which lead to the best performance showing in Table 4, are as follows: for 10 base classifiers 1, 10, and 90%, respectively, for 20 base classifiers 3, 20, and 85%, respectively, for 30 base classifiers 5, 10, and 50%, respectively, for 40 base classifiers 1, 10, and 75%, respectively, and finally, for 50 base classifiers 3, 3, and 25%, respectively.

Overall, the number of base classifiers does not seem to impact effectiveness much, so using a small number (e.g., 10) is recommended in order to avoid high costs of TBM.

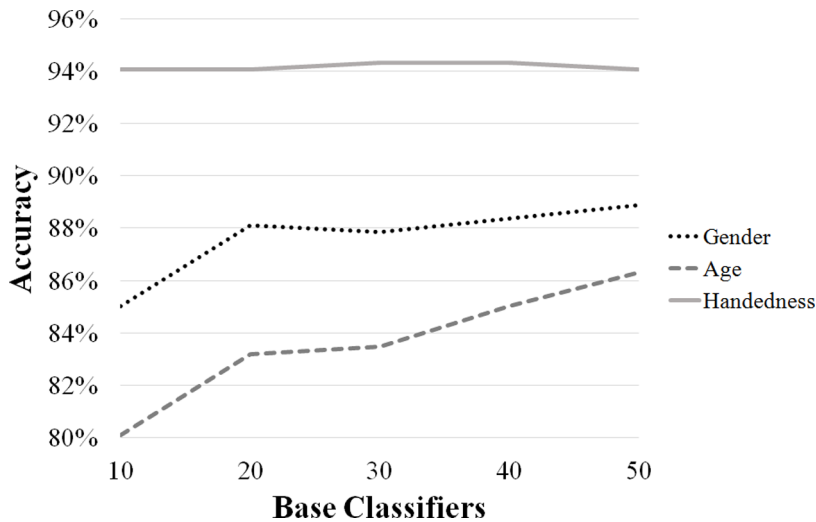
*Table 4. Performance of rotation forest in the handedness classification problem for different numbers of base classifiers*

Base Classifiers	Acc.	TBM (secs)	F1	AUC
10	94.1%	1.19	0.930	0.967
20	94.1%	1.47	0.931	0.939
30	94.3%	3.70	0.934	0.964
40	94.3%	5.25	0.935	0.959
50	94.1%	9.43	0.933	0.958

## Discussion of the Results

The summary results of the experiments in terms of accuracy in the three classification problems examined are presented in Figure 1.

Figure 1. Accuracy in the three classification problems over different number of base classifiers



As shown in Figure 1 and Tables 2, 3, and 4, the accuracy in each of the three classification problems far exceeds the baseline. The baseline can be defined as the percentage of instances of the dominant class in the dataset. Thus, the baseline in gender classification case is considered 52.4%, while the highest accuracy that measured is 88.9%. In the case of age classification, the baseline is considered 33.3%, while the highest accuracy is 86.3%. Finally, in the handedness classification the baseline is 88.6% and the highest accuracy is 94.3%.

Regarding the improvement of accuracy in relation to the increase in the number of base classifiers, different behavior is observed in each of the three cases. In the search for user handedness, the accuracy does not seem to increase with the increase of the number of base classifiers and it seems that the 10 C4.5 decision trees are more than enough to achieve the highest accuracy. In the search for gender, there is a significant improvement of accuracy as the number of base classifiers increases from 10 to 20, and then, with the further increase in the number of C4.5 decision trees the accuracy improves at a much lower rate. Finally, in the search of the age group that a user belongs to, there is also a significant improvement of accuracy between 10 and 20 base classifiers, but in contrast to gender classification there is also a significant improvement between 30 and 50 base classifiers. So, a higher accuracy in gender and age classification may be achieved by using more base classifiers. However, this goes beyond the scope of the present study, which is to use the rotation forest for the first time in experiments with keystroke dynamics data and to check whether it has promising results. Therefore, the search for the



highest possible accuracy in gender and age classification is shifted to a possible extension of this research.

Also, another important parameter of the operation of the rotation forest during user profiling, which must be taken into account, is the training time required. In gender classification the highest accuracy is 88.9% and is achieved with a training time of 23.09 seconds. With a tradeoff of 0.8% in accuracy the training time is reduced by about 60%, to 9.33 seconds. In age classification the highest accuracy observed is 86.3% which is achieved with training time 29.92 seconds, while with a tradeoff of 1.3%, approximately 55% less time (13.60 seconds) is required. Finally, in the handedness classification the almost highest accuracy, 94.1%, is achieved with training time 1.19 seconds.

The execution of the experiments showed a correlation between the training time and the number of base classifiers. The more base classifiers the longer the training time, since for each additional base classifier an additional iteration is performed in the algorithm. Also, the training time is affected by the percentage of removed instances. The higher the removal rate, the shorter the training time, since a smaller training set is created.

## **SOLUTIONS AND RECOMMENDATIONS**

The rotation forest seems quite promising in creating the profile of completely unknown users utilizing data from the way they type. However, there are two other issues that need to be decided.

Firstly, the keystroke dynamics features to be used in the process. Due to the large number of available features the Chi-Square feature selection procedure was followed and all those features that presented a non-zero Chi-Square value were used. Usually, using more features leads to higher accuracy, but it also leads to systems with longer training time. In the present study it was not tested whether the use of only some of the features that showed non-zero Chi-Square value would lead to the creation of a system with similar, or even higher, accuracy and shorter training time. Also, it was not tested whether the use of features with zero Chi-Square value would lead to the creation of systems with similar, or even shorter, training time and higher accuracy. Those two experiments go beyond the objectives of the present study. In any case, choosing the number of features that will be used, as well as exactly which features will be used, is a decision that depends on how accurate the system must be and how fast it must work.

Secondly, a second tradeoff is again between accuracy and training time, but this time it concerns the number of base classifiers that will be in the ensemble. As stated in the “Discussion of the Results” subsection, it is possible to choose to create

an accurate system that runs at a specific time, or a less accurate system that runs faster. The decision to be made will take into account which is the most important criterion, accuracy or training time.

## **FUTURE RESEARCH DIRECTIONS**

In the present study it was shown that rotation forest can be used in user classification using keystroke dynamics data with high accuracy. This research can be extended in different directions.

Firstly, in terms of the performance of the rotation forest, as mentioned above, experiments with a larger number of base classifiers should be conducted in order to check the performance of the model and find the highest accuracy that can be achieved, especially in gender and age classification problems. Moreover, something that has also been mentioned is conducting additional experiments that will use a different set of features than what the Chi-Square feature selection procedure indicated. In addition, although the C4.5 decision tree is proposed to be the base classifier, experiments could be conducted using other base classifiers, such as other decision trees, Bayesian classifiers, k-nearest neighbors, or others.

Secondly, in terms of user attribution, other user characteristics could be sought, such as educational level, mother tongue, height (which is related to the length of the fingers), computer experience, etc. For this purpose, additional data should be collected from a significant number of users so that each defined class is adequately represented. In this direction of extending the research, and if several user characteristics that can be detected with high accuracy are included, the ultimate goal would be to create a system that uses keystroke dynamics features to create the profile of an unknown user so that it can either be used in the case a digital forensics investigation, or to facilitate the use of computers and Internet services, or to be used to protect unsuspecting users. Clearly, there are some issues that need to be addressed. These are, on the one hand, the consent of the users for the recording of their typing, and on the other hand, the way in which the recording will be done in order to avoid the disclosure of sensitive and personal data. One suggestion is to integrate the keylogging application into the operating systems and to perform the extraction of keystroke dynamics features locally. These features will be sent to dedicated servers which will be responsible for evaluating user characteristics, but also for updating databases with labeled data. In this way, data from users whose identity cannot be revealed will be shared, as well as will be used only after the user's choice, except of course in cases of prosecutorial intervention.

Third, since a very large percentage of users connect to the Internet through mobile devices, the research should be extended to seek the characteristics of users

of these devices. For this reason, a suitable keylogger should be developed and data from typing on smartphones and tablets should be collected. Although there are differences in the study of typing between portable and non-portable devices, such as the fact that additional features can be utilized, like the pressure exerted on the touch screen, the methodology to be followed will be similar.

Finally, as far as keystroke dynamics studies are concerned, a possible extension is to look for a correlation between user characteristics and how the keys are used depending on their position on the keyboard. That is, for example, to consider whether left-handed users use the left part of the keyboard differently from right-handed users, in terms of the time intervals required to use a key, a digram, etc., or, if males use the keyboard numpad differently than females. Such an extension of the research may lead to the revelation of some hidden patterns that will develop user profiling.

The present research, with the help of machine learning, seems to be able to develop into an important tool of cybersecurity.

## **CONCLUSION**

Rotation forest is an ensemble machine learning model that uses a number of base classifiers, usually decision trees, and can perform classification or regression. Although it was proposed 15 years ago and has shown very good performance in various problems, it has not been used to date in user classification with keystroke dynamics data. In this work, user profiling is attempted with data coming from the way users type and with the help of the rotation forest which uses the C4.5 decision tree as the base classifier. Specifically, the gender, the age, and the handedness of unknown Internet users are predicted, and the highest accuracy achieved was 88.9%, 86.3%, and 94.3%, respectively. The results show that the use of rotation forest in keystroke dynamics classification problems is very promising and can be the basis of a machine learning system that will serve as a cybersecurity tool.

## **REFERENCES**

Antal, M., & Nemes, G. (2016). Gender recognition from mobile biometric data. In *Proceedings of 11th International Symposium on Applied Computational Intelligence and Informatics* (pp. 243-248). Timisoara, Romania: IEEE. 10.1109/SACI.2016.7507379

- Buker, A., & Vinciarelli, A. (2021). Who is typing? Automatic gender recognition from interactive textual chats using typing behaviour. In A. E. Hassanien, A. Darwish, S. M. Abd El-Kader, & D. A. Alboaneen (Eds.), *Enabling Machine Learning Applications in Data Science. Algorithms for Intelligent Systems* (pp. 3–15). Springer. doi:10.1007/978-981-33-6129-4\_1
- Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal*, 20(1), 131–148. doi:10.1177/1536867X20909693
- Degtereva, V., Gladkova, S., Makarova, O., & Melkostupov, E. (2020). Forming a mechanism for preventing the violations in cyberspace at the time of digitalization: Common cyber threats and ways to escape them. In *Proceedings of the International Scientific Conference - Digital Transformation on Manufacturing, Infrastructure and Service* (article no.: 55, pp. 1–6). ACM. 10.1145/3446434.3446468
- Earl, S., Campbell, J., & Buckley, O. (2021). Identifying soft biometric features from a combination of keystroke and mouse dynamics. In M. Zallio, C. Raymundo Ibañez, & J. H. Hernandez (Eds.), *Advances in Human Factors in Robots, Unmanned Systems and Cybersecurity. Lecture Notes in Networks and Systems* (Vol. 268, pp. 184–190). Springer. doi:10.1007/978-3-030-79997-7\_23
- Fairhurst, M., & Da Costa-Abreu, M. (2011). Using keystroke dynamics for gender identification in social network environment. In *Proceedings of 4th International Conference on Imaging for Crime Detection and Prevention 2011* (pp. 1-6). London, UK. IET. 10.1049/ic.2011.0124
- Giot, R., Dorizzi, B., & Rosenberger, C. (2015). A review on the public benchmark databases for static keystroke dynamics. *Computers & Security*, 55, 46–61. doi:10.1016/j.cose.2015.06.008
- Hossain, M. S., & Haberfeld, C. (2020). Touch behavior based age estimation toward enhancing child safety. In *Proceedings of 2020 IEEE International Joint Conference on Biometrics* (pp. 1-8). IEEE. 10.1109/IJCB48548.2020.9304913
- Idrus, S. Z. S., Cherrier, E., Rosenberger, C., & Bours, P. (2014). Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords. *Computers & Security*, 45, 147–155. doi:10.1016/j.cose.2014.05.008
- Krysowski, E., & Tremewan, J. (2020). Why does anonymity make us misbehave: Different norms or less compliance? *Economic Inquiry*, 59(2), 776–789. doi:10.1111/ecin.12955

### **User Profiling Using Keystroke Dynamics and Rotation Forest**

Lam, K. H., Meijer, K. A., Loonstra, F. C., Coerver, E. M. E., Twose, J., Redeman, E., Moraal, B., Barkhof, F., de Groot, V., Uitdehaag, B. M. J., & Killestein, J. (2020). Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis. *Multiple Sclerosis Journal*, 27(9), 1421–1431. doi:10.1177/1352458520968797 PMID:33150823

Lee, H., Hwang, J. Y., Kim, D. I., Lee, S., Lee, S. H., & Shin, J. S. (2018). Understanding keystroke dynamics for smartphone users authentication and keystroke dynamics on smartphones built-in motion sensors. *Security and Communication Networks*, 2018, 2567463. Advance online publication. doi:10.1155/2018/2567463

Mastoras, R. E., Iakovakis, D., Hadjidimitriou, S., Charisis, V., Kassie, S., Alsaadi, T., Khandoker, A., & Hadjileontiadis, L. J. (2019). Touchscreen typing pattern analysis for remote detection of the depressive tendency. *Scientific Reports*, 9(1), 13414. doi:10.103841598-019-50002-9 PMID:31527640

Nitzburg, G. C., & Farber, B. A. (2019). Patterns of utilization and a case illustration of an interactive text-based psychotherapy delivery system. *Journal of Clinical Psychology*, 75(2), 247–259. doi:10.1002/jclp.22718 PMID:30628062

Papadatou-Pastou, M., Ntolka, E., Schmitz, J., Martin, M., Munafo, M. R., Ocklenburg, S., & Paracchini, S. (2020). Human handedness: A meta-analysis. *Psychological Bulletin*, 146(6), 481–524. doi:10.1037/bul0000229 PMID:32237881

Rachburee, N., & Punlumjeak, W. (2015). A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. In *Proceedings of 7th International Conference on Information Technology and Electrical Engineering* (pp. 420-424). IEEE. 10.1109/ICITEED.2015.7408983

Raul, N., Shankarmani, R., & Joshi, P. (2020). A comprehensive review of keystroke dynamics-based authentication mechanism. In A. Khanna, D. Gupta, S. Bhattacharyya, V. Snasel, J. Platos, & A. Hassanien (Eds.), *International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing* (vol. 1059, pp. 149-162). Springer. 10.1007/978-981-15-0324-5\_13

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10), 1619–1630. doi:10.1109/TPAMI.2006.211 PMID:16986543

Roy, S., Roy, U., & Sinha, D. (2018). Identifying soft biometric traits through typing pattern on touchscreen phone. In J. Mandal & D. Sinha (Eds.), *Social Transformation – Digital Way. Communications in Computer and Information Science* (Vol. 836, pp. 546–561). Springer. doi:10.1007/978-981-13-1343-1\_46

Tsimperidis, I., Peikos, G., & Arampatzis, A. (2021). Classifying users through keystroke dynamics. In T. Chadjipadelis, B. Lausen, A. Markos, T. R. Lee, A. Montanari, & R. Nugent (Eds.), *Data Analysis and Rationality in a Complex World. Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 311-319). Springer. doi:10.1007/978-3-030-60104-1\_34

Tsimperidis, I., Rostami, S., Wilson, K., & Katos, V. (2021). User attribution through keystroke dynamics-based author age estimation. In B. Ghita & S. Shiaeles (Eds.), *Selected Papers from the 12th International Networking Conference. Lecture Notes in Networks and Systems* (vol. 180, pp. 47-61). Springer. 10.1007/978-3-030-64758-2\_4

Udandaraao, V., Agrawal, M., Kumar, R., & Shah, R. R. (2020). On the inference of soft biometrics from typing patterns collected in a multi-device environment. In *Proceedings of 2020 IEEE Sixth International Conference on Multimedia Big Data* (pp. 76-85), IEEE. 10.1109/BigMM50055.2020.00021

Ulinskasa, M., Damaseviciusa, R., Maskeliunasa, R., & Wozniak, M. (2018). Recognition of human daytime fatigue using keystroke data. *Procedia Computer Science, 130*, 947–952. doi:10.1016/j.procs.2018.04.094

Uzun, Y., Bicakci, K., & Uzunay, Y. (2015). *Could we distinguish child users from adults using keystroke dynamics?* <https://arxiv.org/abs/1511.05672>

Van Balen, N., Ball, C. T., & Wang, H. (2016). A Behavioral biometrics based approach to online gender classification. In R. Deng, J. Weng, K. Ren, & V. Yegneswaran (Eds.), *12th International Conference on Security and Privacy in Communication Networks* (pp. 475-495). Springer International Publishing. 10.1007/978-3-319-59608-2\_27

Vesel, C., Rashidisabet, H., Zulueta, J., Stange, J. P., Duffecy, J., Hussain, F., Piscitello, A., Bark, J., Langenecker, S. A., Young, S., Mounts, E., Omberg, L., Nelson, P. C., Moore, R. C., Koziol, D., Bourne, K., Bennett, C. C., Ajilore, O., Demos, A. P., & Leow, A. (2020). Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A BiAffect iOS study. *Journal of the American Medical Informatics Association, 27*(7), 1007–1018. doi:10.1093/jamia/ocaa057 PMID:32467973

Wong, T. T., & Yang, N. Y. (2017). Dependency analysis of accuracy estimates in k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering, 29*(11), 2417–2427. doi:10.1109/TKDE.2017.2740926

Woods, K., Kegelmeyer, W. P. J., & Bowyer, K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(4), 405–410. doi:10.1109/34.588027

## ADDITIONAL READING

Brizan, D. G., Goodkind, A., Koch, P., Balagani, K., Phoha, V. V., & Rosenberg, A. (2015). Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies*, 82, 57–68. doi:10.1016/j.ijhcs.2015.04.005

Pentel, A. (2017). High precision handedness detection based on short input keystroke dynamics. In *Proceedings of 8th International Conference on Information, Intelligence, Systems & Applications* (pp. 1-5). IEEE. 10.1109/IISA.2017.8316380

Plank, B. (2018). Predicting authorship and author traits from keystroke dynamics. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media* (pp. 98-104). The COLING 2016 Organizing Committee. 10.18653/v1/W18-1113

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10), 1619–1630. doi:10.1109/TPAMI.2006.211 PMID:16986543

Thaseen, I. S., Kumar, C. A., & Ahmad, A. (2019). Integrated intrusion detection model using Chi-Square feature selection and ensemble of classifiers. *Arabian Journal for Science and Engineering*, 44(4), 3357–3368. doi:10.1007/13369-018-3507-5

Tsimperidis, I., & Arampatzis, A. (2020). The keyboard knows about you: Revealing user characteristics via keystroke dynamics. *International Journal of Technoethics*, 11(2), 34–51. doi:10.4018/IJT.2020070103

Tsimperidis, I., Arampatzis, A., & Karakos, A. (2018). Keystroke dynamics features for gender recognition. *Digital Investigation*, 24, 4–10. doi:10.1016/j.diin.2018.01.018

## KEY TERMS AND DEFINITIONS

**Chi-Square Test:** The procedure used to examine the differences between categorical variables.

**Digital Forensics:** The process of uncovering and interpreting electronic data.

**Digram Latency:** The time elapsed between the pressing or releasing of a key and the pressing or releasing of the next key.

**Feature Selection:** The process of reducing the number of input variables when developing a predictive model.

**Keystroke Duration:** The time elapsed between the pressing and the releasing of a key. In the literature it is also found as dwell time, or hold time, or press hold, or key press time.

**Keystroke Dynamics:** The way a user uses a keyboard, physical or virtual.

**User Profiling:** The process of identifying some characteristics of a user.