The Keyboard Knows About You: Revealing User Characteristics via Keystroke Dynamics

Ioannis Tsimperidis
Democritus University of Thrace, Greece

Avi Arampatzis
Democritus University of Thrace, Greece

ABSTRACT

One of the causes of several problems on the Internet, such as financial fraud, cyber-bullying, and seduction of minors, is the complete anonymity that a malicious user can maintain. Most methods that have been proposed to remove this anonymity are either intrusive, or violate privacy, or expensive. This paper proposes the recognition of certain characteristics of an unknown user through keystroke dynamics, which is the way a person is typing. The evaluation of the method consists of three stages: the acquisition of keystroke dynamics data from 118 volunteers during the daily use of their devices, the extraction and selection of keystroke dynamics features based on their information gain, and the testing of user characteristics recognition by training five well-known machine learning models. Experimental results show that it is possible to identify the gender, the age group, the handedness, and the educational level of an unknown user with high accuracy.

Keywords: Keystroke Dynamics, User Characteristic Classification, Data Mining, Feature Selection, Information Gain, Digital Forensics

INTRODUCTION

Today there are more than 4 billion Internet users in the world who use online services in order to communicate, entertain, educate, work, etc. The way we communicate over the Internet with someone else differs radically from the way we do it in person. Most of the time we do not see the face of our interlocutor, nor his/her expressions, we do not hear his/her voice, nor the way its tone changes. The stimuli that used to give us information about who our interlocutor is and what his/her intentions are, have ceased to exist. In addition, we have to consider that often a user is talking to someone completely unknown and that kids and teenagers participate in these conversations, especially in social networks. It is easily understood that these lurk many dangers, such as financial fraud, seduction of minors, anonymous threats, etc. In addition, it raises the question of how ethical it is for someone to take advantage of this particularity of communication and to conceal his/her identity from his/her interlocutor.

According to definition of "Technoethics" from the work of Alim and Khalid (2019), technology (apart from being part of social development) causes changes in lifestyle and as a result many ethical considerations have to be addressed. Much of these considerations are about the individuals, and more specifically the ethical questions that are exacerbated by the ways in which technology extends or curtailed their power. Consequently, the limitation imposed on a computer user to know some things

about the person talking, through a messaging application for example, is an issue to be considered in the light of "Technoethics". Just as it would in a face-to-face conversation, or even a telephone conversation, where everyone would receive information about their interlocutor, consequently modifying their attitude accordingly, it would be fair to do so where this information cannot be obtained.

One solution to the aforementioned problem is to know some characteristics of the user we are talking to, but without violating his/her privacy, such as the user's gender, age, educational level, and so on. There are several proposals for achieving these, such as that of Cheung and She (2017), who tried to recognize the gender of users from the images generated by their mobile devices and shared in social networks. The gender of the user can also be predicted from multimodal data as demonstrated by Estruch *et al.* (2017) using a corpus containing text data, but also image and location information, coming from three different social networks from users of three different cities, achieving accuracy of 91.3%. A method for recognizing the age of users that exploits sociolinguistically-based and content-related text features is proposed by Simaki *et al.* (2016), while similarly, Arroju *et al.* (2015) try to determine the gender and age of Twitter users based on the contents of their tweets. Although in most cases the target characteristic is the gender and/or age of users, there are also methods in the literature trying to discover other characteristics, such as the work of Seneviratne *et al.* (2014) where the authors attempt to determine religion, relationship status, spoken languages, and countries of interest of unknown users from snapshots of apps installed on their smartphones.

All the aforementioned approaches rely on machine-learning models and they showed some limitations in finding the characteristics of a user who tries to hide. For example, some of the proposed methods require special equipment, such as special cameras or keyboards, or can only be applied if the target user has an account in some social network, while some others use features derived from certain phrases, words, N-grams, and characters of a language, and therefore are incapable of dealing with the variety of languages and jargon in today's Internet. Consequently, in cases where some characteristics of a user who is attempting to maintain anonymity are sought, in a forensic investigation for example, probably none of the above methods would be effective or even applicable. In contrary, methods based on keystroke dynamics features are free from such limitations. This is because the only device keystroke dynamics methods need is the common QWERTY keyboard, which is an integral part of desktops and laptops, as well as tablets and smartphones, in its virtual form. Furthermore, keystroke dynamics methods seem to be language independent since the features derive mainly from how the user uses the keyboard rather than the words he/she writes in a specific language. Finally, data can be collected non-obtrusively or even without interfering with users' ongoing work or consent, preserving also their privacy content-wise.

Keystroke dynamics can be described briefly as the way a user handles the keyboard. This is interpreted by how long a user needs to use two, three, or more consecutive keys, which of the duplicate keys ("Shift", "Ctrl", "Alt") prefers, how often makes typing errors and how he corrects them (*i.e.* the use of "Backspace", "Insert", "Delete", *etc.*), how often does pauses while typing and what their duration is, etc. The possibility of identifying an individual through keystroke dynamics is being studied for over 40 years now, and a variety of systems (Ali *et al.*, 2016) have been proposed to replace the traditional authentication scheme using passwords.

The keystroke dynamics features used can be categorized into temporal and non-temporal. The most commonly used temporal features are keystroke duration (the time elapsed from pressing to releasing a key) and digram latency (the time elapsed between two consecutive keystrokes). The latter can be expressed in four different ways, which are down-down, up-up, down-up, and up-down digram latency (El-Abed *et al.*, 2014), depending on whether the hit or release time of a keystroke is considered. Other temporal features, such as those related to trigrams, tetragrams and n-grams, are reported by Sim and Janakiraman (2007). Common non-temporal features are those related to typing speed, error rate, pauses during typing, specific keys usage, etc. (Alsultan *et al.*, 2017).

The present study is not yet another work on user authentication, as is the case with most studies on keystroke dynamics, but an attempt to identify some inherent or acquired characteristics of unknown

users, namely gender, age, handedness, and educational level, which aims to provide information to a computer user about who their interlocutor is.

The rest of the paper is organized as follows. Firstly, in "Background" section, the related works in user classification through keystroke dynamics are listed. Secondly, in "Method" section, we describe the data acquisition, the keystroke dynamics features extraction, and the feature selection procedures. Thirdly, in "Experiments and Results" section, the classification results obtained by using five well-known machine learning models, *i.e.* the support vector machine (SVM), the simple logistic (SL), the Bayes classifier (NB), the Bayesian network classifier (BNC) (Jing *et al.*, 2008), and the radial basis function network (RBFN), are presented. The results are then discussed, in "Discussion" section, and finally the paper is concluded, in "Conclusion" section.

BACKGROUND

Although most research into keystroke dynamics has as its object user authentication, there are some published papers, especially over the last five years, in user classification. Once again, the characteristic sought in most cases is users' gender, followed by age. For example, Li et al. (2019) studied the possibility of identifying the gender of a user participated in online chatting. They collected data from communication among 45 volunteers (35 males and 10 females), and ensured that the training dataset was gender-balanced. Keystroke durations and all types of digram latencies were used as keystroke dynamics features, but they extracted only from the 20 most commonly used digrams in English language. To identify the gender of the author of a message they proposed a system which combines Random Forest, a score-level fusion, and a majority voting mechanism, and managed to achieve a correct prediction of 76%. In another work, Plank (2018) used an already existing dataset, which after filtering to remove data that did not meet certain criteria, consisted of typing sessions from 121 participants (53 females and 68 males). Her purpose was to identify who was the author of a text among a group of known users, or to identify the gender and age of the author. With the help of the most commonly used keystroke dynamics features she achieved an F-score of 0.635 on gender recognition, and 0.733 on age group recognition between 2 classes.

Buriro *et al.* (2016) investigated the possibility to estimate user characteristics on smart mobile devices, namely gender, age, and handedness, when users type a PIN/password from 4 to 16 digits. They collected their data from 150 volunteers on a specific device and defined 3 age groups, teenagers (<20), adults (≥20 and <60), and senior users (≥60). They extracted temporal keystroke dynamics features and used Naïve Bayes, SVM, Random Forest, MLP, and Deep Neural Network for classification. The best results came from Random Forest (RF), which had an accuracy of 82.8% in gender classification, 87.9% in age classification, and 95.5% in handedness classification. Random Forest was also the most successful classifier among 7 others, in the work of Roy *et al.* (2017). They conducted their study to protect kids from unknown threats coming from the Internet and therefore divided users into two classes, kids and adults. They used three fixed text datasets from 11 to 14 keystrokes, two created in desktop computers and one in a touch screen device, and exploited keystroke durations, down-down, up-down, and up-up digram latencies. Finally, using an Ant Colony Optimization (ACO) technique they achieved an accuracy of 92.2%.

Studies with more age classes are those of Tsimperidis *et al.* (2017) who divided the users into 4 groups (18-25, 26-35, 36-45, 46+), and Pentel (2018) into 6 groups (<16, 16-19, 20-29, 30-39, 40-49, 50+). The former study used 120 down-down digram latencies as features, and with a dataset of 239 log files presented 66.1% accuracy coming from MLP combined with a boosting algorithm, while the latter study with data from more than 7,000 users, each of which was recorded for about 320 keystrokes, and 134 keystroke dynamics features in total reached 61.6% accuracy using Random Forest.

Handedness is a human characteristic that has been extensively researched in terms of economy, sociology, biology, criminology, etc. (Fagard *et al.*, 2017). In the field of user classification through keystroke dynamics, Brizan *et al.* (2015) collected data from 329 users who answered 10-12 questions in a closed environment. They used keystroke dynamics features, namely keystroke durations, digram

latencies, mean times associated with using common and rare keys, etc., but also stylometric and language production features. The experimental results showed an F-score of 0.223 for the left-hand class, with a baseline of 0.1, using LogitBoost, Naïve Bayes, SVM, and Simple Logistic. Shen *et al.* (2016) exploited a dataset created by 51 users (43 right-handers and 8 left-handers) who typed an 11 character password several times, and extracted keystroke durations, down-down, and up-down digram latencies as features. They used their own weighted Random Forest and achieved an accuracy of 87.75%. Another approach is that of Pentel (2017) who collected data from 504 users (403 right-handers and 101 left-handers) through an electronic questionnaire using JavaScript code. The entire dataset consisted of only 43 keystrokes by each user, on average, but using Naïve Bayes, Logistic regression, Simple Logistic, SVM, Nearest Neighbor, C4.5, and Random Forest managed to present high performance. Initially, keystroke durations and digram latencies were used and the F-score was 0.643, and then six more features based on location on keyboard were added and the F-score rose to 0.995, with the baseline being 0.5 due to balancing of dataset. Similarly, in the work of Shute et al. (2017) 65 volunteers (54 right-handers and 11 left-handers) were recorded in the same laptop while typed 2-7 particular long texts each, producing 421 log files. The authors split the keyboard into six segments, namely "upper", "middle", and "lower", which each has a "left" and "right", and then fed the features, which were keystroke durations only, in C4.5, Neural Network, and Random Forest classifiers resulting in an accuracy of 94.5%.

User classification studies based on how users use the keyboard are quite rare. There may be no other published work in seeking age and handedness of unknown users other than those mentioned. In fact, we have not found any paper referring to user classification according to educational level, which is one of the main focuses of our work. This makes it interesting to conduct for the first time a study on whether (and to what extent) it is possible to classify users on the basis of this acquired characteristic, the results of which can be used in a similar way as those of inherent characteristics.

METHOD

Our methodology consists of three consecutive phases. In the first phase, we collected free-text data from volunteers who agreed to participate in the experiment of extracting real-life keystroke dynamics features, in order to create a dataset to be used for conducting experiments of the present research. In the second phase, we ran a feature selection algorithm to sort the features according to their contained information, in order to identify those features that contain the most information, so as to ignore a multitude of them that do not contribute much to the performance of user classification procedure, thus reducing the run time of experiments. In the third phase, the gender, the age, the handedness, and the educational level of an unknown user are sought by training and hyperparameter-tuning five well-known machine learning algorithms, namely SVM, Simple Logistic, Naïve Bayes, Bayesian Network, and RBFN, in order to find a model that can distinguish users in the best way in terms of their characteristics.

Keystroke Dynamics Dataset

Keystrokes dynamics datasets can be created by recording users either in fixed- or in free-text. The term "fixed-text" refers to the typing of a specific text usually in some closed environment, while "free-text" indicates the recording of a volunteer during the typical daily use of his/her computer. On the one hand, by using fixed-text, researchers can focus on particular features while the sensitive data of the user remain secure. On the other hand, using free-text may reveal features which contain more information. In this work, the free-text approach is followed as it integrates with the subject's regular typing activities better and is less intrusive.

To create a suitable dataset, a free text keylogger named "IRecU", which can be installed on any Microsoft Windows-based devices, was designed and developed. In each of the volunteers who participated in this project, "IRecU" was installed on their personal computer and it was possible to record their typing at anytime, anywhere they wanted to work, and using any application, gathering data from thousands keystrokes, in order to get the best possible approximation of the actual use of the computer. In contrast, the creation of other free-text keystroke dynamics datasets in the literature was

done with volunteers being recorded on a specific device, or at a specific time (*e.g.* some sessions), or in a specific location (*e.g.* office or lab), or in a specific application (*e.g.* browser), or by collecting data from a few keystrokes. In addition, to protect volunteers from disclosing their passwords and personal messages to a third person, firstly a signed statement was given to them that the data would only be used for this research, secondly it was given an option to use "IRecU" whenever they want, and thirdly an opportunity was given to review (but not modify) the recorded data so they can decide whether to share the log file or not.

There are two issues to consider here. First, as with any experiment where the subject knows that he/she is being observed, it is likely that he/she will change his/her behavior resulting to recorded data which may not be representative. Of course, the recording method followed simulates much more the daily use of the computer by the users than the other methods presented in "Background" section, and is therefore considered an improved approach. Second, the ability of volunteers to use "IRecU" whenever they wanted could have skewed the dataset. This is due the fact that some volunteers could use the keylogger constantly, while others in specific applications. However, following this policy, by not specifying certain times of the day and specific applications where volunteers will be recorded, the non-harassing process and the privacy of the participants are best ensured. In this way, the data is obtained by using computers that approximate normal use as much as possible.

The recording of volunteers was completed in two periods. The first one had a duration of 10 months, from 20/02/2014 to 27/12/2014, where 75 volunteers returned 248 log files, and the second had a duration of 8.5 months, from 24/10/2017 to 09/07/2018, where 43 volunteers returned 139 log files, forming a dataset of 387 log files from 118 users (*i.e.* almost 3.3 files per user). In each file there are some metadata, such as the gender, age group, dominant hand, educational level, mother tongue, etc. of the recorded user, while keystrokes were written in records of the form:

78,#2018-03-19#,45743645,"dn"
79,#2018-03-19#,45743769,"dn"
78,#2018-03-19#,45743785,"up"
79,#2018-03-19#,45743879,"up"
96,#2018-03-19#,45849163,"dn"
96,#2018-03-19#,45849226,"up"

In each record, which is a user's action on the keyboard, there are four fields separated by commas. The first field represents the virtual key code of the key used (from 1 to 255), the second indicates the date the action took place in the yyyy-mm-dd format, the third is the elapsed time since the beginning of that day (12:00am) in milliseconds (from 0 to 86399999), and the forth is the action, "dn" for key-press and "up" for key-release. The log files varied in size from 170 KB to 271 KB and contained data from 2,800 to 4,500 keystrokes.

Demographics of the dataset that are of interest to this research are shown in Table 1. As it can be seen, the dataset is unbalanced in each of the characteristics being studied. However, it is evident that with regards to gender there are almost the same number of male and female volunteers and log files, while with regards to age and educational level each class is adequately represented. With regards to handedness, the dataset is as unbalanced as it should, since the right-handers/left-handers worldwide ratio is approximately 9 to 1 (Cavanagha, 2016).

Table 1. Number of volunteers and log files per gender, age, dominant hand, and educational level

Characteristic	**Class**	**Volunteers**		**Log Files**	
		#	**%**	**#**	**%**
Gender	Male	61	51.7	203	52.4
	Female	57	48.3	184	47.6

Age	18-25	31	26.2	96	24.8
	26-35	37	31.4	129	33.3
	36-45	37	31.4	117	30.2
	46+	13	11.0	45	11.7
Handedness	Right-handers	105	89.0	343	88.6
	Left-handers	10	8.5	35	9.0
	Ambidextrous	3	2.5	9	2.4
Educational Level (According UNESCO)	ISCED-3	21	17.8	62	16.0
	ISCED-4	7	5.9	23	6.0
	ISCED-5	21	17.8	74	19.1
	ISCED-6	36	30.5	120	31.0
	ISCED-7-8	33	28.0	108	27.9

Other demographics of the volunteers are that out of 118, 104 are Greek native speakers, 8 are Turkish native speakers, 5 are English native speakers, and 1 is Bulgarian native speaker.

Feature Extraction and Feature Selection

As described in the Introduction of this paper, there are hundreds of thousands of keystroke dynamics features. In order to keep the complexity low, we considered the most frequently-used features, namely the keystroke durations and down-down digram latencies. The duration of keystrokes is calculated from the subtraction of milliseconds that correspond to the "up" action minus the ms that correspond to the "dn" action, for the same key. Similarly, the down-down digram latency results from the subtraction of ms of a "dn" action minus the ms of the previous "dn" action. For the feature extraction we developed a software application, named "ISqueezeU", which reads the text files created by "IRecU" and calculates the average values of keystroke durations or down-down digram latencies. In order to deal with data sparsity, only the keys that have at least 5 appearances and the digrams with at least 3 appearances have been taken into account, while for the other ones the values were marked as unknown.

Although we chose to extract a small part of the available keystroke dynamics features, their number is n^2+n, with n being the number of keyboard keys. This means that more than 10,000 features were extracted, which is a large number that can lead to systems with high time complexity. Therefore, a feature selection procedure is needed.

Of the thousands of features, there must be selected those which are most capable of distinguishing users according to the studied characteristics, namely gender, age, handedness, and educational level. A method to do this is by calculating the information gain (*IG*) of each feature *f*, which is the measure that illustrates the ability of that feature to reduce the entropy of a system *x*. It is expressed as:

$$\mathrm{IG}(x,\mathrm{f}) = H(x) - H(x|\mathrm{f}) \qquad (1)$$

The entropy $H(x)$ of the system x is given by:

$$H(x) = -\sum_{i=1}^{m} P(x_i)\ln P(x_i) \qquad (2)$$

In Equation (2), m is the length of vector x, which in the classification problem is the number of classes, and $P(x_i)$ is the probability of class x_i. In our case we have 2, 4, 3, and 5 classes for gender, age, handedness, and education level, respectively. With the probabilities of each class for each classification problem being those shown in Table 1, the entropy of the system is 0.692 in the gender classification

problem, 1.325 in the age classification problem, 0.413 in the handedness classification problem, and 1.497 in the educational level problem.

The term $H(x|f)$ is calculated by splitting the dataset into groups according to the value of the particular feature f. Then, the entropy of each group is calculated and $H(x|f)$ is given by:

$$H(x|\mathrm{f}) = \frac{1}{N}\sum_{j=1}^{k} n_j H(x_j) \qquad (3)$$

where N is the number of instances of the initial dataset, k is the number of groups that the initial dataset was split, n_j is the number of instances of the j-th group, and $H(x_j)$ is the entropy of the j-th group, which can be calculated from Equation (2).

This procedure is also described in the work of Dash *et al.* (2013), and if applied to every extracted feature in our classification problems, then a list with the amount of information that every feature carries will emerge. In Table 2, the first 15 features are ranked with the highest *IG* for gender, age, handedness, and educational level classification problems, where each of them is represented by the virtual key code of the keys that compose it. So, one number indicates keystroke duration and two numbers indicate down-down digram latency.

Table 2. Keystroke dynamics features with the highest IG in gender, age, handedness, and educational level classification

#	Gender			Age			Handedness			Educational Level		
	Feat.	**Keys**	***IG***	**Feat.**	**Keys**	***IG***	**Feat.**	**Keys**	***IG***	**Feat.**	**Keys**	***IG***
1	68	D	0.0586	69	E	0.1457	79	O	0.0832	83	S	0.1431
2	80-65	P-A	0.0569	65-32	A-(space)	0.1377	65	A	0.0769	32	(space)	0.1301
3	73-78	I-N	0.0550	79	O	0.1006	82-65	R-A	0.0703	76	L	0.1149
4	77-65	M-A	0.0532	65	A	0.0802	84-65	T-A	0.0656	76-186	L-;:	0.1050
5	78-65	N-A	0.0515	68	D	0.0791	69	E	0.0592	186	;:	0.0932
6	65	A	0.0504	32	(space)	0.0781	65-84	A-T	0.0506	80	P	0.0904
7	75-65	K-A	0.0457	39	(right arrow)	0.0746	82	R	0.0493	89	Y	0.0879
8	77-79	M-O	0.0428	87	W	0.741	71	G	0.0489	77	M	0.0857
9	87	W	0.0422	83	S	0.0721	83-84	S-T	0.0418	85	U	0.0847
10	79-78	O-N	0.0419	89	Y	0.0689	186	;:	0.0392	84	T	0.0829
11	78-79	N-O	0.0411	86	V	0.0659	66	B	0.0386	75-186	K-;:	0.0800
12	84-79	T-O	0.0407	84-79	T-O	0.0637	76-69	L-E	0.0382	79	O	0.0799
13	79-77	O-M	0.0404	87-32	W-(space)	0.0620	32-65	(space)-A	0.0371	71	G	0.0793
14	76-69	L-E	0.0402	70	F	0.0618	65-32	A-(space)	0.0361	73-75	I-K	0.0793
15	69-73	E-I	0.0397	88	X	0.0592	84	T	0.0322	72	H	0.0769

Some observations that can be made from Table 2 are: a) keystroke durations seem to play more important role than digram latencies in age and educational level classification problems, while digram latencies are more significant in gender classification problem, b) the “A”, “M”, “N”, and “O” keys (along with the digrams in which they participate) carry significant amount of information in case of gender classification, c) the keys “A” and “T” (along with the digrams in which they participate) carry significant amount of information in case of handedness classification, and d) in case of educational level

classifications the letter keys located on the right side of the keyboard (“O”, “P”, “K”, “L”, “;:”, “N”, and “M”) appear among the top positions in the ranking.

Experimental Procedure and Validation of Models

The feature selection procedure indicated 514, 690, 246, and 727 features with non-zero information gain on the gender, age, handedness, and educational level classification problems, respectively. Since we try to create systems with high precision in predicting user characteristics, we decided to take advantage of any feature that carries some information, according to the analysis we made, and thus all those with non-zero information gain were used.

Several machine learning models were tested, such as Random Forest, C4.5, k-Nearest Neighbors, Random Tree, OneR, MLP, etc., which presented low accuracy, even bellow the baselines, and/or too long training times. The five models which presented the best performance in terms of accuracy and time complexity were Support Vector Machine (SVM), Simple Logistic (SL), Naïve Bayes (NB), Bayesian Network classifier (BNC), and Radial Basis Function Network (RBFN). Therefore, the results of these models will be presented.

The purpose of model validation is to ensure that the implementations of the models are correct and work as they should. There are many techniques that can be utilized to verify a model and several of them were adopted to validate the five models used in this work.

Firstly, to assess the performance of the five models fairly, we use the well-known 10-folds cross-validation, which divides the data into 10 disjoint parts, uses 9 of them for training and the remaining one for testing, in a round-robin fashion (Jung, 2018). In our case where we have 387 log files, each part in which the dataset is divided consists of 38 or 39 files. Also, the vast majority of volunteers delivered 3-4 log files. With these numbers 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.

Secondly, to evaluate the effectiveness of the feature selection procedure we additionally use F-score, 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 F-score is a measurement of how balanced is the prediction between classes. For example, assume two cases of a system for our handedness classification problem. In the first case, the system predicts all users as right-handed. The accuracy is almost 89%, but it is obvious that the system is not working properly. In the second case, the system correctly predicts the dominant hand of users 8 out of 9 instances, for right-handed, left-handed, and ambidextrous. The accuracy is again 89%, but this system is more reliable. This greater reliability is reflected in the F-score, where in the latter case is higher.

Finally, to assess the ranking ability of the classifiers we use the area under the ROC curve (AUC) or ROC index (Obuchowski & Bullen, 2018). 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 - precision. The ROC curve is limited to the interval [0, 1] in both dimensions, thus AUC varies between 0 and 1.

EXPERIMENTS AND RESULTS

For each classification problem contemplated in this paper and for each of the five mentioned models, a large number of experiments were conducted to find the values of classifiers’ hyperparameters that lead to the best performance, in terms of accuracy (Acc.), which is the percentage of correctly classified instances, the time complexity (TBM---Time to Build Model), which is the CPU time has taken to build model, the F-score (F1), which is the harmonic mean of recall and precision, and the ROC index (AUC), which is the area under the ROC (Receiver Operating Characteristic) curve.

Gender Classification

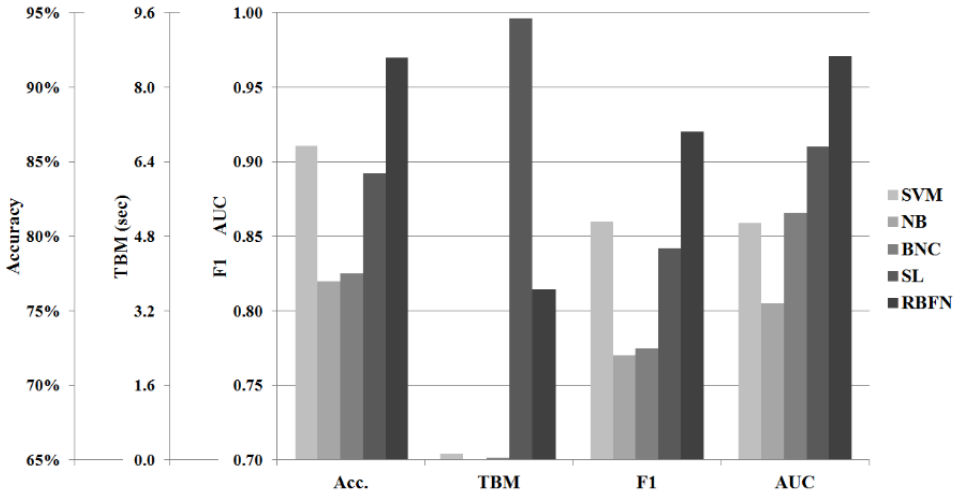
After carrying out the experiments, the results obtained for the gender classification problem are those shown in Table 3.

Table 3. Performance of the five models in the gender classification problem

Model	Acc.	TBM	F1	AUC
SVM	86.1%	0.13	0.860	0.859
NB	77.0%	0.02	0.770	0.805
BNC	77.5%	0.05	0.775	0.866
SL	84.2%	9.48	0.842	0.910
RBFN	92.0%	3.67	0.920	0.971

The best performance shown in Table 3 was achieved for SVM having a Polykernel (polynomial kernel) as kernel type and value 1.2 for the C parameter. For BNC it was achieved having the SimpleEstimator to estimate the conditional probability tables of a Bayes network, 0.001 as initial count on each value for estimating the probability tables, K2 algorithm for searching network structures, and 1 as the maximum number of parents of each node in Bayes network. For SL it was achieved having 500 as the maximum number of iterations for LogitBoost, 180 as the last iteration of LogitBoost if no new error minimum has been reached, and 100% for weight trimming. For RBFN it was achieved having 130 clusters for K-means, and 2.8 minimum standard deviation for the clusters. Figure 1 visualizes the results of Table 3.

Figure 1. Performance of SVM, NB, BNC, SL, and RBFN in gender classification problem



The comparison of the statistical values of the gender classification methods reported in section "Background" is shown in Table 4, where the baseline values of accuracy and F-score were derived from the percentage of male and female log files in the training dataset. The larger of the two percentages was considered the baseline value. Where "---" is shown, it means that there is no information for this statistical value. As mentioned earlier, gender is the most studied characteristic in classifying users through keystroke dynamics, and there is clearly more work than listed. However, a representative part is provided for comparison purposes.

Table 4. Performance comparison of gender classification methods

Method	Acc.	Acc. Baseline	F1	F1 Baseline

Li *et al.* (2019)	76.0%	50.0%	---	---
Plank (2018)	---	---	0.635	0.562
Buriro *et al.* (2016)	82.8%	50.0%	---	---
Our best run, RBFN	92.0%	52.4%	0.920	0.524

Age Classification

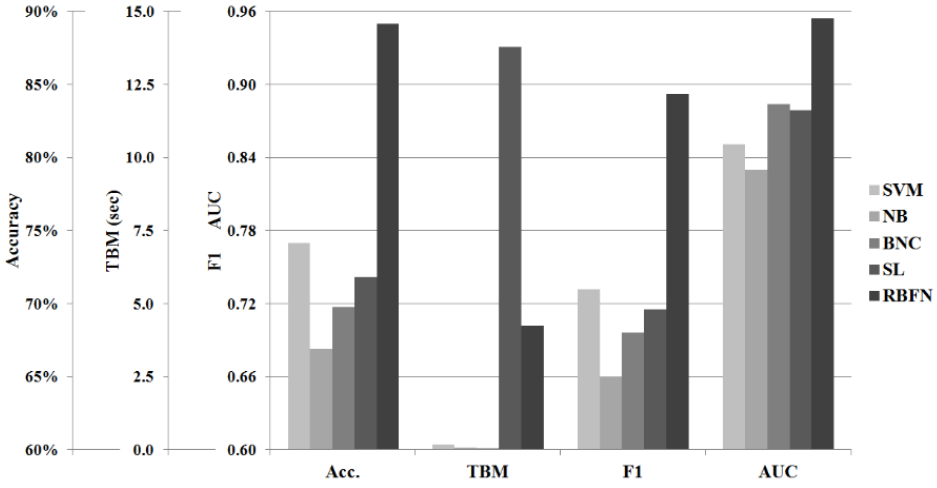
The results after hyperparameter tuning for the age classification problem are shown in Table 5.

Table 5. Performance of the five models in the age classification problem

Model	Acc.	TBM	F1	AUC
SVM	74.2%	0.19	0.732	0.851
NB	66.9%	0.08	0.660	0.830
BNC	69.8%	0.06	0.696	0.884
SL	71.8%	13.78	0.715	0.879
RBFN	89.2%	4.25	0.892	0.954

The best performance shown in Table 5 was achieved for SVM having Polykernel and C parameter equal to 0.5. For BNC it was achieved having SimpleEstimator, 0.02 as initial count on each value for estimating the probability tables, K2 algorithm, and 1 as the maximum number of parents. For SL it was achieved having 500 as the maximum number of iterations, 100 as the last iteration, and 100% for weight trimming. For RBFN it was achieved having 110 clusters, and 1.2 minimum standard deviation. Figure 2 graphically presents the statistical values of the five models in Table 5.

Figure 2. Performance of SVM, NB, BNC, SL, and RBFN in age classification problem



A comparison of the methods performance for classifying users based on their age, as they reported in the section "Background", is presented in Table 6, where the baseline accuracy value is given by the most likely random choice, that is, the class with the most samples. Where no quota information of classes is given in the corresponding paper, the dataset is considered to be balanced. Moreover, the values denoted with "---" indicate that no reference is made to them in the corresponding paper.

Table 6. Performance comparison of age classification methods

Method	# of Classes	Acc.	Acc. Baseline	F1	AUC

Buriro *et al.* (2016)	3	87.9%	33.3%	---	---
Roy *et al.* (2017)	2	92.2%	50.0%	---	---
Tsimperidis *et al.* (2017)	4	66.1%	25.0%	0.658	---
Pentel (2018)	6	61.6%	16.7%	0.620	0.880
Our best run, RBFN	4	89.2%	33.3%	0.892	0.954

There may be no direct comparison of the results presented in Table 6 because each study used a different number of classes, divided users differently into age groups, and exploited datasets produced in a variety of ways. In this respect, the comparison is left at the discretion of the reader. What we can claim about the superiority of our own method is that, in addition to the high accuracy rate relative to the number of classes, the data was obtained from the most realistic use of computer keyboards, compared to all other methods.

Handedness Classification

In the problem of handedness classification the results are shown in Table 7.

Table 7. Performance of the five models in the handedness classification problem

Model	Acc.	TBM	F1	AUC
SVM	94.6%	0.09	0.940	0.811
NB	90.7%	0.05	0.896	0.637
BNC	95.6%	0.02	0.954	0.959
SL	95.9%	0.91	0.956	0.957
RBFN	97.2%	1.09	0.973	0.964

In Table 7, SVM used Polykernel and value 0.8 for the C parameter, BNC used SimpleEstimator, 0.05 initial count for estimating the probability tables, K2 algorithm, and one parent at most for each node, SL used 500 maximum iterations for LogitBoost, 60 iteration as the limit to stop LogitBoost, and 90% for weight trimming, RBFN used 70 clusters for K-means with 0.5 minimum standard deviation for them. These results are also shown graphically in Figure 3.

Figure 3. Performance of SVM, NB, BNC, SL, and RBFN in handedness classification problem

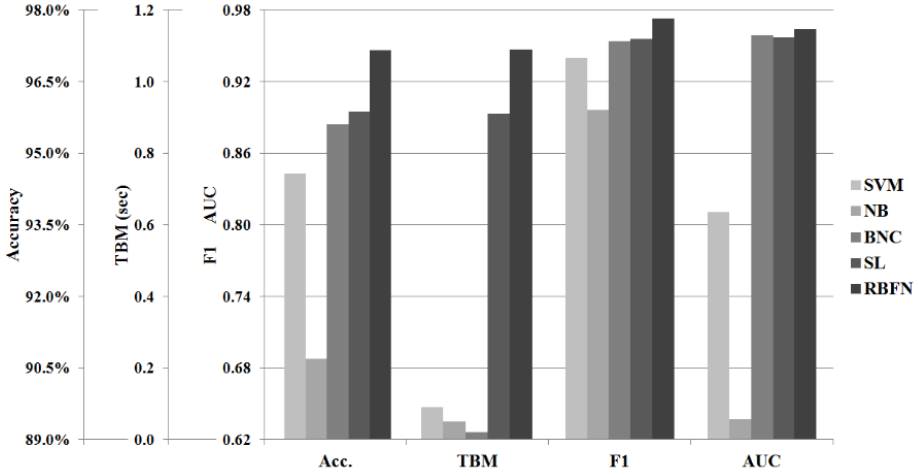


Table 8 compares the user classification methods based on the dominant hand mentioned in section "Background".

Table 8. Performance comparison of handedness classification methods

Method	# of Classes	Acc.	Acc. Baseline	F1	F1 Baseline
Shen *et al.* (2016)	2	87.8%	84.3%	---	---
Buriro *et al.* (2016)	2	95.5%	90.7%	---	---
Pentel (2017)	2	---	---	0.995	0.500
Shute *et al.* (2017)	2	94.5%	83.0%	---	---
Our best, RBFN	3	97.2%	88.6%	0.973	0.886

The baseline values were selected to be the percentage of instances came from right-handed users, unless otherwise stated in the relevant paper. This is because, in the case of handedness classification due to the highly unbalanced dataset, the random prediction of the dominant hand of an unknown user would be the "right hand". The missing values denoted with "---".

As it can be seen from Table 8, the qualitative difference between datasets used by the different methods has led to a variety of baselines. Therefore, everyone can set their own criteria for deciding which method is superior. For example, is it preferable to start from a baseline of 83.0% and achieve an accuracy of 94.5%, or start from a baseline of 88.6% and achieve an accuracy of 97.2%? Also, there is a great difference in the size of the datasets used by each method. Some methods have used records that consist of a few keystrokes, while we used thousands of keystrokes. Again, one can argue that using more data leads to more reliable systems, while others can argue that some systems achieve high accuracy rates with a small number of keystrokes. Leaving the reader again to decide, we stress the fact that we seem to be the first to use three classes, dedicating one of them to ambidextrous, and that we acquired the keystroke dynamics data in more realistic environments.

Educational Level Classification

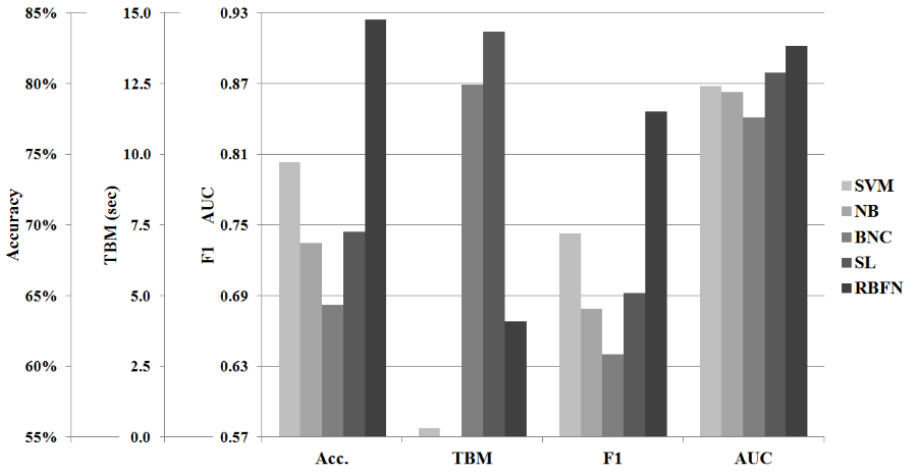
The best results of SVM, NB, BNC, SL, and RBFN for the educational level classification problem are shown in Table 9.

Table 9. Performance of the five models in the educational level classification problem

Model	Acc.	TBM	F1	AUC
SVM	74.4%	0.31	0.743	0.868
NB	68.7%	0.03	0.679	0.863
BNC	64.3%	12.45	0.640	0.841
SL	69.5%	14.34	0.692	0.879
RBFN	84.5%	4.08	0.846	0.902

The best performances shown in Table 9 were derived with the following settings. SVM: Polykernel and C=0.6. BNC: SimpleEstimator, initial count=0.25, K2 algorithm, and parent for each node=5 at most. SL: max iterations for LogitBoost=500, last iteration for LogitBoost if no new min error=80, and weight trimming=100%. RBFN: clusters for K-means=70 and min std dev=0.85. Figure 4 shows the comparison of models in educational level classification problem.

Figure 4. Performance of SVM, NB, BNC, SL, and RBFN in educational level classification problem



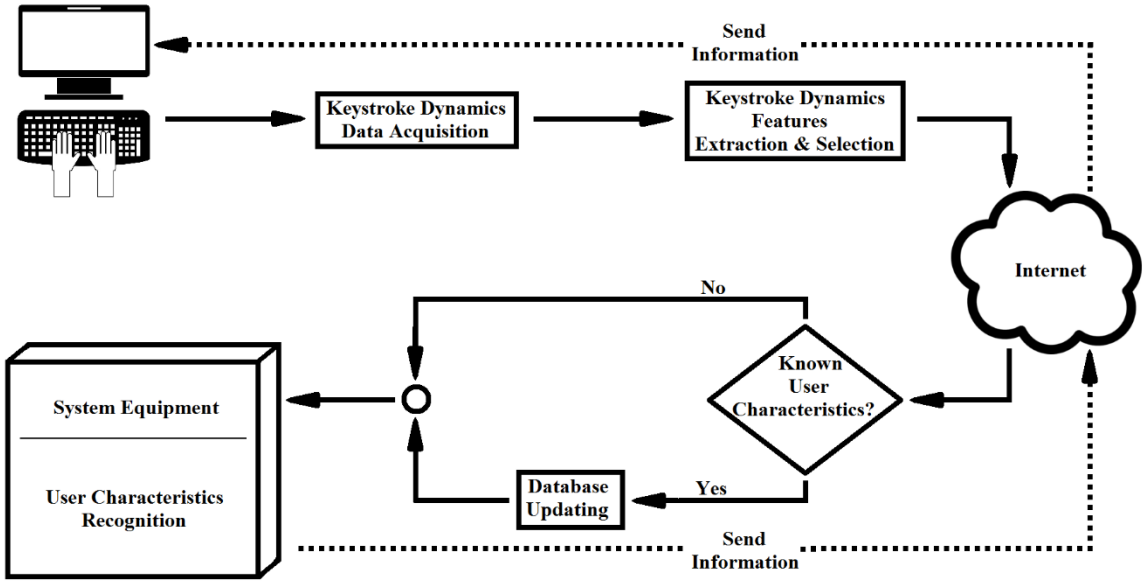
As it was stated before, there is no published work which classifies computer users according to their educational level using keystroke dynamics features, so we cannot compare our results with others.

DISCUSSION

From the experimental results shown in Figures 1 to 4, it can be seen that RBFN outperforms in terms of accuracy, F-score, and area under the ROC curve all other models in every classification problem examined in this work. SVM displays the second best accuracy, except for handedness classification, where SL and BNC proved to be more accurate. The same pattern is shown in F-score, with the SVM having the second highest value behind the RBFN, except for handedness classification. In regards to the area under the ROC curve, a differentiation is observed, where the second best value, behind RBFN, is presented by SL in gender and educational level classification, and by BNC in age and handedness classification. However, the fastest models proved to be NB and BNC, but they have a disadvantage in accuracy and F-score in almost every case. RBFN and SL have the longest training time, but the values shown in Tables 3, 5, 7, and 9 are not prohibitive for their use as they are (without condensation method, reducing the dimensionality, etc.). In conclusion, it seems that the RBFN model is the most suitable for user classification according to some inherent or acquired characteristics, mainly because it correctly predicts the gender, age group, handedness, and educational level of an unknown user with 92.0%, 89.2%, 97.2%, and 84.5%, respectively. Indeed, these high percentages are achieved with a few hundred keystroke dynamics features and at a time that does not exceed a few seconds of model training, even with the computing power of a personal computer. Of course, in order to verify all of the above, new phases of volunteers' recording should be followed to enlarge the existing dataset and repeat the experiments with more data.

Another observation that can be made is that, as mentioned in the "Keystroke Dynamics Dataset" section, data acquisition in our method was done through a process that attempted to approach as much as possible the daily use of computers by users, unlike other methods limited to a small number of keystrokes, a short recording time, or recording in a specific environment. Consequently, the results presented are a strong indication of feasibility of creating systems that can predict characteristics of unknown users only by the way they use the keyboard. Such systems will operate as shown in Figure 5. When users type, the way they use their keyboard is recorded, and then the desired keystroke dynamics features are extracted. This process is executed locally on the user's computer/device, so no one else has access to these sensitive data. Keystroke dynamics features are transferred to the system server, where user characteristics are recognized. In case that these features come from a user with known characteristics, they are used to update the database. Finally, the system provides information on gender, age, handedness, educational level, and maybe other user characteristics.

Figure 5. The operation of a user characteristics recognition system employing keystroke dynamics



A proposal for implementation of such a system, which is also suggested in similar works (Tsimperidis *et al.*, 2018), is its embedment to operating systems. In this way, once the keystroke dynamics data collected, and once the desired features extracted, they are sent to a dedicated server which is responsible for deciding on the characteristics of the user. There are two points worth paying attention to. First, by sending keystroke dynamics features instead of the data itself, it is not possible to mine sensitive or personal information such as passwords, personal messages, etc. In addition, for more security, keystroke dynamics data recorded locally, as well as keystroke dynamics features that are transmitted over the Internet, could be encrypted. Second, knowing that patterns are not as prominent as the typing speed, but rather they are quite hidden, it would be very difficult for a user to modify his/her typing rhythm so as to conceal his/her characteristics, especially when the proposed system dynamically adapts its parameters triggered by the availability of new training data. In order to prove the above allegations, further research will be needed, such as the development of an application that captures online keystroke dynamics data, extracts the appropriate features, and test of how well the aforementioned results hold up. However, this goes beyond the objectives of this study.

Some applications of such a system are first to inform unsuspecting users about the characteristics of their interlocutor so as to avoid misleading them if a malicious user tries to exploit them by counterfeiting some of his/her characteristics, *e.g.* age in seduction of minors. Second, to provide valuable information in case of forensic investigation, for example when an electronic crime has been committed and some characteristics of the offender are recognized, thereby excluding from the proceedings persons whose gender, age, handedness, and educational level do not match with findings. Third, to facilitate the user, since by recognizing some of his/her characteristics, it will be possible to automatically fill in fields in forms, to personalize the advertising addressed to him/her, and to suggest websites and groups on Internet of his/her interest.

CONCLUSION

Often, full anonymity on the Internet can make it difficult for users to access useful services, or even worse, be the advantage of malicious users. Existing methods that achieve user characteristics recognition require specific data, such as facial images, or are intrusive, or violate privacy by accessing, for example, texts written by users. On the contrary, keystroke dynamics provide a non-intrusive low-cost method using data coming only from the way users use the keyboard.

This study presents a process in which the most suitable keystroke dynamics features are selected to identify the gender, the age, the handedness, and the educational level of an unknown user. To accomplish the objective, a new keystroke dynamic dataset was created from recording users during the daily usage of their devices, and 387 log files were collected. The information gain of each feature was then calculated and they were ranked according to the reduction of entropy of the system. The experimental results showed that it is possible to create quite reliable systems that can recognize the aforementioned four characteristics of an unknown Internet user with accuracy of 92.0%, 89.2%, 97.2%, and 84.5%, respectively, using only a few hundred features and with a short time of model training.

Having the ability to recognize some characteristics of an unknown user who types a certain piece of text has significant value in digital forensics, targeted advertisement, and facilitating users. But beyond all that, it is an ethical issue to enable someone to know who their interlocutor really is. However, we note that the deployment of such a system must be in accordance with the current legal and regulatory framework, as the unauthorized analysis of keystrokes may entail hidden privacy violations, which might involve sensitive personal information (e.g., in accordance to the EU legislation).

In conclusion, the present work recognizes the problem of lack of stimulation during Internet communication, through messaging applications for example, which would had a person during a face-to-face conversation. To avoid any problems that this may cause, such as cheating unsuspecting users from malicious ones, a method is proposed for the first time (according to our knowledge) in the literature, which exploits keystroke dynamics features to give computer users the necessary information about their interlocutor. The method is based on the fact that it utilizes data related to how a user types, rather than what he/she types, thus ensuring that personal or sensitive data will not be exposed, while safeguarding everyone's ethic right to know a few things about the person he/she is talking to.

There are many directions in which this study can be extended. One of these is to conduct experiments with keystroke dynamics features except keystroke durations and diagram latencies, which are the most commonly used features. Such features may relate to user typing habits, such as the number and duration of typing pauses, the change of typing rate, the percentage use of duplicate keys, and so on. Another possible extension is the further acquisition of keystroke dynamics data, mainly from users of different mother tongues, in order to examine the claim that the keystroke dynamics methods are language independent. Finally, the proposed method can be approached, as well, with Dempster-Shafer's theory of evidence (Su *et al.*, 2015), considering each keystroke dynamics feature as a "view" which must be combined with other "views" to produce the final outcome.

REFERENCES

Ali, M. D., Monaco, J. V., Tappert, C. C., & Qiu, M. (2016). Keystroke biometric systems for user authentication. *Journal of Signal Processing Systems, 86*(3), 175-190.

Alim, S., & Khalid, S. (2019). Support for cyberbullying victims and actors: A content analysis of Facebook groups fighting against cyberbullying. *International Journal of Technoethics, 10*(2), 35-56.

Alsultan, A., Warwick, K., & Wei, H. (2017). Nonconventional keystroke dynamics for user authentication. *Pattern Recognition Letters, 89*, 53-59.

Arroju, M., Hassan, A., & Farnadi, G. (2015). Age, gender and personality recognition using tweets in a multilingual setting. In *Proceedings of 6th Conference and Labs of the Evaluation Forum: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Toulouse, France (pp. 23-31).

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.

Buriro, A., Akhtar, Z., Crispo, B., & Del Frari F. (2016). Age, gender and operating-hand estimation on smart mobile devices. In *Proceedings of 2016 International Conference of the Biometrics Special Interest Group*, Darmstadt, Germany (pp. 273-280).

Cavanagha, T., Berbesquea, J. C., Wood, B., & Marlowe, F. (2016). Hadza handedness: Lateralized behaviors in a contemporary hunter–gatherer population. *Evolution and Human Behavior 37*(3), 202-209.

Cheung, M., & She, J. (2017). An analytic system for user gender identification through user shared images. *ACM Transactions on Multimedia Computing, Communications, and Applications, 13*(3), 30:1-30:20.

Dash, S. K., Dash, A. P., Dehuri, S., & Sung-Bae Cho, S. (2013). Feature selection for designing a novel differential evolution trained radial basis function network for classification. *International Journal of Applied Metaheuristic Computing 4*(1), 32-49.

El-Abed, M., Dafer, M., & El Khayat, R. (2014). RHU Keystroke: A mobile-based benchmark for keystroke dynamics systems. In *Proceedings of 2014 International Carnahan Conference on Security Technology*, Rome, Italy (pp. 1-4).

Estruch, C. P., Palacios, R. P., & Rosso, P. (2017). Learning multimodal gender profile using neural networks. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Varna, Bulgaria (pp. 577-582).

Fagard, J., Margules, S., Lopez, C., Granjon, L., & Huet, V. (2017). How should we test infant handedness? *Laterality: Asymmetries of Body, Brain and Cognition, 22*(3), 294-312.

Jing, Y., Pavlović, V., & Rehg, J. M. (2008). Boosted Bayesian network classifiers. *Machine Learning, 73*(2), 155-184.

Jung, Y. (2018). Multiple predicting K-fold cross-validation for model selection. *Journal of Nonparametric Statistics, 30*(1), 197-215.

Li, G., Borj, P. R., Bergeron, L., & Bours, P. (2019). Exploring keystroke dynamics and stylometry features for gender prediction on chat data. In *Proceeding of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics*, Opatija, Croatia (pp. 1233-1238).

Obuchowski, N. A., & Bullen J. A. (2018). Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine & Biology 63*(7), 07TR01.

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

Pentel, A. (2018). Predicting user age by keystroke dynamics. In R. Silhavy (Ed.), *Artificial intelligence and algorithms in intelligent systems* (pp. 336-343). Switzerland: Springer International Publishing.

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*, New Orleans, Louisiana, USA (pp. 98-104).

Roy, S., Roy, R., & Sinha, D. D. (2017). ACO-Random forest approach to protect the kids from Internet threats through keystroke. *International Journal of Engineering and Technology, 9*(3S), 279-285.

Seneviratne, S., Seneviratne, A., Mohapatra, P., & Mahanti, A. (2014). Predicting user traits from a snapshot of apps installed on a smartphone. *ACM SIGMOBILE Mobile Computing and Communications Review, 18*(2), 1-8.

Shen, C., Xu, H., Wang, H., & Guan, X. (2016). Handedness recognition through keystroke-typing behavior in computer forensics analysis. In *Proceedings of 2016 IEEE Trustcom/BigDataSE/ISPA*, Tianjin, China (pp. 1054-1060).

Shute, S., Ko, R. K. L., & Chaisiri, S. (2017). Attribution using keyboard row based behavioural biometrics for handedness recognition. In *Proceedings of 2017 IEEE Trustcom/BigDataSE/ICESS*, Sydney, NSW, Australia (pp. 1131-1138).

Sim, T., & Janakiraman, R. (2007). Are digraphs good for free-text keystroke dynamics?. In *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA (pp. 1-6).

Simaki, V., Mporas, I., & Megalooikonomou, V. (2016). Age identification of Twitter users: Classification methods and sociolinguistic analysis. In *Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics*, Konya, Turkey (pp. 385-395).

Su, X., Mahadevan, S., Xu, P., & Deng, Y. (2015). Handling of dependence in Dempster–Shafer Theory. *International Journal of Intelligent Systems, 30*(4), 441-467.

Tsimperidis, I., Arampatzis, A., & Karakos, A. (2018). Keystroke dynamics features for gender recognition. *Digital Investigation, 24*, 4-10.

Tsimperidis, I., Rostami, S., & Katos, V. (2017). Age detection through keystroke dynamics from user authentication failures. *International Journal of Digital Crime and Forensics, 9*(1), 1-16.