# User attribution through keystroke dynamics-based author age estimation

Ioannis Tsimperidis [1], Shahin Rostami [2], Kevin Wilson [2], and Vasilios Katos [2]

[1] Democritus University of Thrace, Komotini, Greece
[2] Bournemouth University, Poole, UK

**Abstract.** Keystroke dynamics analysis has often been used in user authentication. In this work, it is used to classify users according to their age. The authors have extended their previous research in which they managed to identify the age group that a user belongs to with an accuracy of 66.1%. The main changes made were the use of a larger dataset, which resulted from a new volunteer recording phase, the exploitation of more keystroke dynamics features, and the use of a procedure for selecting those features that can best distinguish users according to their age. Five machine learning models were used for the classification, and their performance in relation to the number of features involved was tested. As a result of these changes in the research method, an improvement in the performance of the proposed system has been achieved. The accuracy of the improved system is 89.7%.

**Keywords:** Keystroke Dynamics Dataset, User Age Classification, Feature Selection, Information Gain, RBFN, AUC, Digital Evidence.

## 1 Introduction

In the original study [1] the authors proposed a system to collect information about an attacker who, by stealing the identity of a legitimate user, had managed to enter a computer system illegally. This information included inherent characteristics, such as gender and age, and also acquired characteristics, such as educational level and computer experience. This information was then used in a forensic investigation to find the guilty party. The research was focused on the task of trying to classify unknown users into age groups, by exploiting data that came from the way a user uses the keyboard. Specifically, by using 120 digram latencies, one of the most widely used keystroke dynamics features, a success rate of 66.1% was achieved in predicting which age group the user belonged to. It is noted that users were divided into four age groups, giving a random prediction rate of 25%.

It is clear that any such system, that can detect the age or other characteristics of a totally unknown user, can be used as an investigation tool in any computer crime. For example, if someone tries to break in to a computer system, tries to mislead an unsuspecting user, or carries out some cyberbullying, it is likely that the malicious user will use a keyboard - in which case, the system can create a profile giving valuable infor-

mation to digital forensics agents. Indeed, the need for such a system, or for something similar, is becoming increasingly urgent due to the amount of cybercrime in the world today [2].

In the previous research, some goals were set for further work, including expanding the available dataset, by recording more users during the daily use of their computers, and using multiple classifiers in parallel, the results from which would be summarised with the use of Dempster-Shafer theory [3]. To achieve these goals, a three step plan was developed. First, a new user recording phase took place; second, more features of the keystroke dynamics were utilized; and third, additional classifiers were used.

The objective of the project has remained the same, which is to recognize certain user traits in the most universal and most economical way possible, whilst ensuring the privacy of the users and minimizing their harassment.

The rest of the paper is organized as follows. In the next section, the relevant literature in this and similar fields is summarised. The following section describes the experimental methods used: namely the data acquisition, the selection of suitable features, and the evaluation procedure. Then the results from the use of five well-known machine learning models, namely support vector machine (SVM), simple logistic (SL), Bayes classifier (NB), Bayesian network classifier (BNC), and radial basis function network (RBFN), are presented. Finally, the paper concludes by listing the conclusions and plans for further research.

## 2 Background

Classifying computer users according to their characteristics could be useful in a variety of applications. For example, in automated translation, due to the fact that many languages have some system of grammatical gender, a more successful translation might result when the gender of the user who is typing is known. Another example is targeted advertising, where characteristics such as gender, age and educational level are important parameters for determining a user's interests. Yet another example is in strengthening user authentication, where the more user characteristics that can be used for comparison, the greater the possibility that the user can be correctly identified. Human-computer interaction applications can also benefit from the ability to recognise user characteristics, as it enables them to modify their interface, and display user specific messages. Because of these and other applications, the classification of computer users has captured the attention of many researchers.

For example, Zhang et al. [4] exploited text features typed by users under different conditions, such as when writing a new message or responding to someone else's message, and tried to classify them according to their gender and age, dividing them into two age groups. With the help of an LSTM network, they managed to achieve a successful prediction rate of 90.8% in gender classification, and 82.3% in age classification. In another paper, Culotta et al. [5] collected data from Twitter user profiles and enriched them with demographic data from an audience measurement company, thus creating their dataset. Using features derived from the association of users with

other users, such as who the user follows, and text features from the tweets that they wrote, and employing a distantly-supervised regression model, they were able to determine the user gender with an F1 value of 0.87. They were also able to determine the nationality, out of four different choices, with an F1 value of 0.81, and also the political orientation, out of two political parties, with an F1 value of 0.74.

One approach to classifying users is to use features extracted from facial pictures, as in the work of Zhang et al. [6], who extracted features from five points of the face, and then cropped the image in different ways. For each resulting image, they used a convolutional neural network to perform gender classification, where they achieved a success rate of about 91.7%, and smile classification, where they achieved a maximum success rate of about 89.3%. Another attempt is that of Chikkala et al. [7], who used data from four facial picture databases and divided users into six age groups. Their method was based on the third order four pixel pattern and achieved a 96% success rate in each of the databases they used, surpassing performance over all other active methods.

Most user classification methods, such as those mentioned above, are based on features that were extracted from users' photos, videos they appear in, text that they typed, websites that they visited, or a profile that they maintain in a social network. Of course, each of these methods serves the purpose for which it was proposed. However, with regard to the search for forensic evidence in cases where a cybercrime has been committed, most, if not all, of these methods are considered inadequate.

There are several reasons for this. One is that a malicious user will usually try to conceal or misrepresent their identity, and so will not perform any attacks through their genuine social network account, Internet service, or computing system, and so there will be no available picture of the attacker. Consequently, methods that perform user classification through facial images, or the examination of website traffic, or through user profile data, cannot be used for this purpose. Another reason is that methods that examine text written by the user, are based on the extraction of the required features from words and other parts of speech, such as digrams and trigrams, from a particular language. This means that these methods have serious limitations when they are used to examine text from a different language.

An alternative approach, which overcomes the aforementioned problems, is to use keystroke dynamics, a field of computer science that studies how users type. The advantage of keystroke dynamics information is that it uses features from the simplest and most common form of communication between Internet users, which is text [8]. Users write emails, send instant messages, make searches in search engines, upload posts, and communicate much more frequently with text than any other method of communication, such as videoconferencing. Another advantage of using keystroke dynamics information is that no special equipment is required for their operation, except for the common QWERTY keyboard. The advantages include the non-disturbance of users as data can be collected during their daily use of the computer without requiring additional actions on their part, and independence from the typed language since the features are not related to words in a particular language.

In the past, keystroke dynamics information has been used primarily to authenticate users in order to replace or enhance user authentication by passwords.

Salem and Obaidat [9] proposed an authentication system for Android mobile devices using temporal features such as digram latencies, and non-temporal features such as on-screen pressure, finger positioning, etc. They created an application for recording volunteer's actions and as well as that data, they also used and existing dataset for their experiments. Various classifiers were tested, and MLP, with an EER of 0.9%, proved to be the one with the best performance. In another paper, Saini et al. [10] attempted to authenticate users of portable devices, regardless of their body posture when they use them. They recorded data from mobile phones with users sitting, walking, or relaxing. In the processing, the random forest and kNN classifiers were used, achieving an optimal EER of 4.3%.

As has already been mentioned, most of the research on keystroke dynamics has focused on authenticating users, with only a small percentage being oriented to other applications, such as Kolakowska's work [11], which attempts to recognize the emotional state of a keyboard user. A small amount of the research is concerned with classifying users according to some of their characteristics, such as the research of Tsimperidis et al. [12], in which the authors collected 242 logfiles from volunteers during daily use of their computers, used a combination of an RBFN classifier and a boosting algorithm, and managed to predict correctly the educational level of a user, among five options, with accuracy of 86.8%. In another paper, Brizan et al. [13] used data from 350 volunteers who typed a short piece of text, extracted textual and keystroke dynamics features, and achieved recognition of a user's mother tongue, gender, and handedness, at rates higher than those of random selection.

Regarding the classification of users according to their age with the help of keystroke dynamics, which is the subject of this research, there are some interesting studies. Buriro et al. [14] tried to investigate the possibility of estimating, among other things, the age of a user who types a PIN/password between 4 and 16 digits in length, on a smart mobile device. They collected their data from 150 volunteers on a specific device and defined 3 age groups. 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 87.9%. Random Forest was also the most successful classifier amongst 7 others, in the work of Roy et al. [15]. They conducted their study to protect young people from unknown threats coming from the Internet and therefore divided users into two classes, children and adults. They used three fixed text datasets from 11 to 14 keystrokes and exploited keystroke durations and digram latencies. Finally, using an Ant Colony Optimization (ACO) technique they achieved an accuracy of 92.2%. Pentel [16] divided the users into 6 groups, gathered data from more than 7,000 users, each of which was recorded for about 320 keystrokes, extracted 134 keystroke dynamics features in total, and reached an accuracy of 61.6% using Random Forest.

It is understood that a number of researches have aimed at classifying users based on one or some of their characteristics taking advantage of various types of data, such as face images and posts on social networks. However, only a very small portion of them use data derived from keystroke dynamics. This research is one of the few in user classification through keystroke dynamics and as far as we know the only one that focuses on the exploitation of results in digital forensics.

# 3 Method

The methodology consists of three consecutive phases. In the first phase, free-text data was collected from volunteers who agreed to participate in the experiment. In the second phase, a feature selection algorithm was used to sort the features according to the information that they contain. In the third phase, an attempt was made to determine the previously unknown age of a user by training and hyperparameter-tuning five well-known machine learning algorithms, namely SVM, Simple Logistic, Naïve Bayes, Bayesian Network, and RBFN.

## 3.1 Keystroke Dynamics Dataset

It was stated that one of the ways of improving the results of the previous research was to extend the available dataset. It was decided that the acquisition of data should be done in a way that interferes as little as possible with the daily use of the computer by the users. For this reason, the keylogger was designed to record actions on the keyboard, in any application, without causing any harassment to the user.

Although the research did not intend to capture the text written by a user, it was technically possible to reconstruct it from the data that was recorded. For this reason, guarantees were given to the volunteers who participated, in order to safeguard their sensitive or personal data, such as passwords, credit card numbers, or messages to third parties. Each volunteer was given a signed consent form by the researchers, stating that the recorded data would be encrypted, remain exclusively in their possession, and would not be shared with others in any way. It also stated that only the keystroke dynamics features would be studied, from which it would not be possible to reconstruct the original text. In addition, volunteers were not only made aware of the potential dangers, they were also given the ability to run the keylogger only when they wanted to, so that they could choose which data was recorded. Finally, they were allowed to oversee the data recorded, and decide at any point in the process whether or not they wished to hand it over to the researchers.

Each keyboard action made by the volunteers was recorded in the logfiles, as shown below:

```
78,#2017-11-14#,56861220,"dn"
78,#2017-11-14#,56861368,"up"
65,#2017-11-14#,56861502,"dn"
73,#2017-11-14#,56861728,"dn"
65,#2017-11-14#,56861742,"up"
73,#2017-11-14#,56861883,"up"
```

Each logfile entry consists of four parts separated by commas and corresponds to a key press or release action. The first part is the virtual key code of the key used (from 1 to 255); the second part, delimited by the sharp character (#), is the date on which the action took place; the third part is the exact time at which the action took place, as an integer denoting the ms that have passed since the beginning of the day (12:00

am); and finally the fourth part shows the type of action, with "dn" representing a key press, and "up" representing a key release.

With these additional measures, and using the software developed for this purpose in the previous study, the researchers conducted a second phase of data collection from volunteers who did not participate in the first phase. The second recording phase lasted 8.5 months, from 24/10/2017 to 09/07/2018, and 43 volunteers were selected, thus increasing the number of participants to 118, so that the demographics of the created dataset reflected those of the world population, such as ensuring that the number of males is approximately equal to the number of females, and that the number of right-handed users is about 90% of the sample [17], and most important for the present study, to have satisfactory representation of all age groups.

Table 1 shows the comparison between the initial dataset (first phase of volunteer recording) and the extended dataset (first and second phases of volunteer recording).

**Table 1.** Comparison between initial and extended dataset.

| | Initial Dataset | | Extended Dataset | |
|---|---|---|---|---|
| Age Group | Number of Files | Percentage | Number of Files | Percentage |
| 18-25 | 32 | 13.4% | 96 | 24.8% |
| 26-35 | 102 | 42.7% | 129 | 33.3% |
| 36-45 | 90 | 37.6% | 117 | 30.3% |
| 46+ | 15 | 6.3% | 45 | 11.6% |
| Total | 239 | | 387 | |

As can be seen from Table 1, the expansion of the dataset led to a more even distribution across the age groups, as the logfiles from “18-25” and “46+” age groups, which were the least common in the initial dataset, increased in number so that their share of the overall dataset almost doubled in percentage terms.

Each of 387 logfiles is between 170 KB and 271 KB in size and contains data relating to between 2,800 and 4,500 keyboard actions. This variation in the size of the logfiles is due to two things: the fact that the keylogger was designed to record data of a certain size in bytes, and therefore, depending on the time of the day the volunteer was recorded, and depending on the keys used, the number of recorded keys could have a difference of ±5%. The other fact is that, as stated in the consent form, no volunteer was obliged to complete the recording process, which sometimes created files of smaller than normal size. Eventually, it was decided that only files exceeding a certain size threshold size would be accepted.

### 3.2 Feature Extraction and Feature Selection

Keystroke dynamics encompass a large number of features, which can be divided into two categories: temporal and non-temporal. The temporal features are the most widely used, and they include keystroke durations and digram latencies. Other features in the same category are the trigram, tetragram, and general n-gram latencies; the dura-

tion of pauses during typing; and the typing rate (words per unit of time). The non-temporal features include the percentage use of duplicate keys (“Shift”, “Ctrl”, digits, etc.); the way in which the typing errors are corrected (“Delete”, “Backspace”); and the time of the day the user is typing.

Much of the research involving keystroke dynamics only makes use of a small number of the available features, with most researchers only using some of the keystroke durations and one or more of the digram latencies (down-down, down-up, up-down, and up-up). In the first phase of this research, the authors made use of 120 down-down digram latencies, which were selected according to their incidence.

In this extension to that phase of the research, the intention is to use more keystroke dynamics features and to evaluate them according to the amount of information that they provide for classifying users according to their age. However, the features examined will include those found in most researches, namely the keystroke durations and down-down digram latencies. There are a large number of these, since $n^2+n$ features can be extracted from a keyboard with $n$ keys. Most companies use the PC keyboard with 104 keys as a de-facto standard and therefore the number of extracted features can be as large as 10,920.

This large number of features can lead to systems with high time complexity and therefore a procedure must be followed to reduce their number. This process, which is called feature selection, must identify those features which are most capable of distinguishing users according to their age. One way of doing this is to calculate 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:

$$IG(x, f) = H(x) - H(x \mid f) \qquad (1)$$

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

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

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 this study there are 4 classes and therefore the entropy of the system is 1.312. The term $H(x|f)$ is calculated by dividing 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:

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

where $N$ is the number of instances of the initial dataset, k is the number of groups that the initial dataset was divided into, $n_j$ 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 Osanaiye et al. [18] and, if applied to every extracted feature in the age classification problem, it will produce a list with

the amount of information that every feature carries. A list of 15 features with the highest *IG* is shown in Table 2, where the keystroke durations are represented with one number (such as "69", the first in the list) and digram latencies are represented with two numbers, separated by a dash (such as "65-32", the second in the list).

**Table 2.** Keystroke dynamics features with the highest IG in age classification.

| # | Feat. | IG | # | Feat. | IG | # | Feat. | IG |
|---|---|---|---|---|---|---|---|---|
| 1 | 69 | 0.1457 | 6 | 32 | 0.0781 | 11 | 86 | 0.0659 |
| 2 | 65-32 | 0.1377 | 7 | 39 | 0.0746 | 12 | 84-79 | 0.0637 |
| 3 | 79 | 0.1006 | 8 | 87 | 0.0741 | 13 | 87-32 | 0.0620 |
| 4 | 65 | 0.0802 | 9 | 83 | 0.0721 | 14 | 70 | 0.0618 |
| 5 | 68 | 0.0791 | 10 | 89 | 0.0689 | 15 | 88 | 0.0592 |

As can be seen in Table 2, keystroke durations appear to play a more important role than digram latencies in user classification based on their age.

### 3.3 Experimental Procedure and Validation of Models

The feature selection procedure used showed that more than 90% of the features extracted contain zero *IG*, so they may be excluded from user classification, resulting in a huge reduction in time complexity with a minimal or no reduction in accuracy.

The other features, those with non-zero *IG*, were all used to predict the unknown age of a user. Various classifiers were tested for this purpose, several of which showed very low success rates, such as Random Forest, C4.5, k-Nearest Neighbors, Random Tree, and OneR, while others had a prohibitively long training time, such as the MLP, which was the classifier used in the previous research of the authors. The five models that presented high accuracy and low time complexity were SVM, SL, NB, BNC, and RBFN, and therefore the experimental process continued with them.

The model validation stage 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.

First, to assess the performance of the models fairly, we use the 10-folds cross-validation method. This divides the data into 10 disjoint parts, uses 9 of them for training and the remaining one for testing, in a round-robin fashion. In this study where the dataset consists of 387 log files, each fold will consist of 38 or 39 files.

Second, in order to evaluate the effectiveness of the feature selection procedure, the F-score was also used as a combined measurement of precision and recall, because accuracy alone cannot give the full picture of the overall performance of a model when classes are imbalanced, and also because the F-score is a measurement of how balanced the prediction is between classes.

Finally, in order to assess the ranking ability of the classifiers, use is made of the receiver operating characteristic (ROC) curve, which shows recall as a function of the probability of a false negative, which is equivalent to 1 – precision. The area under

the ROC curve (AUC) or ROC index [19] was used. The ROC curve is limited to the interval [0, 1] in both dimensions, thus the AUC varies between 0 and 1.

## 4 Experiments and Results

For each of the five models (SVM, SL, NB, BNC, and RBFN) several experiments were conducted to find the classifier parameters that implement the system with the optimal performance for different sets of features. The first criterion was the performance with the highest accuracy (Acc.), with the second being the one with the lowest time complexity (TBM - Time to Build Model), followed by the highest Area Under the ROC Curve (AUC) and the highest F-score (F1).

Experiments were done with various sets of features in order to evaluate the performance of these models. These involved using different numbers of keystroke dynamics features, starting with the first 100 features according to their IG value and finishing with 700 features, in steps of 100.

The best performance of SVM for different number of features, along with the optimal C value, is shown in Table 3.

**Table 3.** The performance of SVM over different number of features

| # of Feats. | Statistical Values | | | | Classifier Parameters | |
|---|---|---|---|---|---|---|
| | Acc. | TBM | AUC | F1 | C | Kernel |
| 100 | 65.9% | 0.06 | 0.796 | 0.642 | 0.8 | Polykernel |
| 200 | 71.1% | 0.12 | 0.826 | 0.706 | 2.0 | Polykernel |
| 300 | 72.9% | 0.11 | 0.845 | 0.722 | 1.0 | Polykernel |
| 400 | 73.6% | 0.17 | 0.853 | 0.732 | 2.0 | Polykernel |
| 500 | 74.4% | 0.25 | 0.861 | 0.741 | 4.0 | Polykernel |
| 600 | 75.5% | 0.27 | 0.864 | 0.752 | 4.0 | Polykernel |
| 700 | 74.2% | 0.19 | 0.851 | 0.732 | 0.5 | Polykernel |

Several conclusions can be drawn from Table 3. First, the accuracy and the F-score, in each different set of features, exceeds the corresponding measures of the same classifier in the previous study, which was 56.5% and 0.545 respectively. Second, as expected, time complexity is too low, even when several features are involved. Third, the polynomial kernel works better than the other kernel types.

Similarly, Table 4 shows the performance of SL and the corresponding optimal values for the last iteration of LogitBoost, over the seven different feature sets, if no new error minimum has been reached.

**Table 4.** The performance of SL over different number of features

| # of Feats. | Statistical Values | | | | Classifier Parameters | |
|---|---|---|---|---|---|---|
| | Acc. | TBM | AUC | F1 | Last Iteration | Weight Trimming |
| 100 | 63.1% | 0.56 | 0.826 | 0.625 | 50 | 95% |

| | | | | | | |
|---|---|---|---|---|---|---|
| 200 | 67.4% | 1.86 | 0.859 | 0.672 | 50 | 100% |
| 300 | 71.1% | 1.58 | 0.874 | 0.706 | 60 | 85% |
| 400 | 72.9% | 5.11 | 0.879 | 0.726 | 150 | 95% |
| 500 | 73.6% | 7.59 | 0.879 | 0.734 | 200 | 95% |
| 600 | 73.1% | 10.61 | 0.879 | 0.727 | 80 | 100% |
| 700 | 71.6% | 9.86 | 0.877 | 0.712 | 50 | 100% |

From Table 4 it follows that the Simple Logistic model shows better accuracy and F-score than the corresponding classifier in the prior study, which were 55.7% and 0.552, respectively.

The results for the NB classifier are in Table 5.

**Table 5.** The performance of NB over different number of features

| # of Feats. | Acc. | TBM | AUC | F1 |
|---|---|---|---|---|
| 100 | 62.3% | 0.03 | 0.829 | 0.620 |
| 200 | 67.2% | 0.06 | 0.835 | 0.664 |
| 300 | 67.7% | 0.03 | 0.837 | 0.670 |
| 400 | 68.2% | 0.06 | 0.839 | 0.673 |
| 500 | 66.9% | 0.06 | 0.835 | 0.660 |
| 600 | 66.9% | 0.02 | 0.831 | 0.660 |
| 700 | 66.9% | 0.02 | 0.830 | 0.660 |

Two findings from Table 5 are that the Naïve Bayes model shows improved accuracy and F-score in each set of features, compared to the previous research, which produced the values 50.2% and 0.488 respectively, and that, as expected, the time complexity of the model is very low.

The best results for BNC are shown in Table 6, which also presents the corresponding optimal initial count on each feature set for estimating the probability tables and the optimal maximum number of parents of each node in Bayes network.

**Table 6.** The performance of BNC over different number of features

| # of Feats. | Statistical Values | | | | Classifier Parameters | |
|---|---|---|---|---|---|---|
| | Acc. | TBM | AUC | F1 | Initial Count | Max Number of Parents |
| 100 | 66.2% | 0.38 | 0.836 | 0.660 | 0.10 | 5 |
| 200 | 67.4% | 0.62 | 0.858 | 0.674 | 0.10 | 3 |
| 300 | 67.7% | 1.36 | 0.865 | 0.676 | 0.20 | 3 |
| 400 | 68.7% | 0.05 | 0.875 | 0.685 | 0.01 | 1 |
| 500 | 69.0% | 0.05 | 0.878 | 0.689 | 0.02 | 1 |
| 600 | 70.0% | 0.06 | 0.886 | 0.699 | 0.01 | 1 |
| 700 | 69.8% | 0.08 | 0.886 | 0.697 | 0.01 | 1 |

Table 6 reveals the seemingly contradictory result that the BNC presents higher time complexity when the number of features involved is smaller, which is due to the different settings of the classifier that led to its best performance in each case. The BNC model was not examined in the previous work and no direct comparison can be made.

Finally, the results from the optimal configuration of RBFN in terms of the number of clusters for K-Means and the minimum standard deviation for the clusters yielding the best performance, are presented in Table 7.

**Table 7.** The performance of RBFN over different number of features

| # of Feats. | Statistical Values | | | | Classifier Parameters | |
|---|---|---|---|---|---|---|
| | Acc. | TBM | AUC | F1 | # of Clusters | Min Std Dev |
| 100 | 82.7% | 1.30 | 0.917 | 0.827 | 130 | 1.1 |
| 200 | 86.6% | 2.31 | 0.942 | 0.866 | 110 | 1.1 |
| 300 | 88.9% | 2.70 | 0.950 | 0.889 | 110 | 1.4 |
| 400 | 89.7% | 3.55 | 0.960 | 0.897 | 120 | 1.2 |
| 500 | 89.2% | 4.22 | 0.949 | 0.891 | 110 | 1.4 |
| 600 | 89.2% | 5.03 | 0.953 | 0.892 | 110 | 1.2 |
| 700 | 89.2% | 5.67 | 0.954 | 0.892 | 110 | 1.2 |

As can be seen from Table 7, the RBFN presents the best performance for each set of features at similar values of the classifier's parameters, namely a value between 110 and 130 for the number of clusters for K-Means, and a value between 1.1 and 1.4 for the minimum standard deviation for the clusters. The RBFN model was also not considered in the previous study.

### 4.1 Evaluation and Comparison of Results

The best performance of each of the examined models is illustrated in Figure 1.

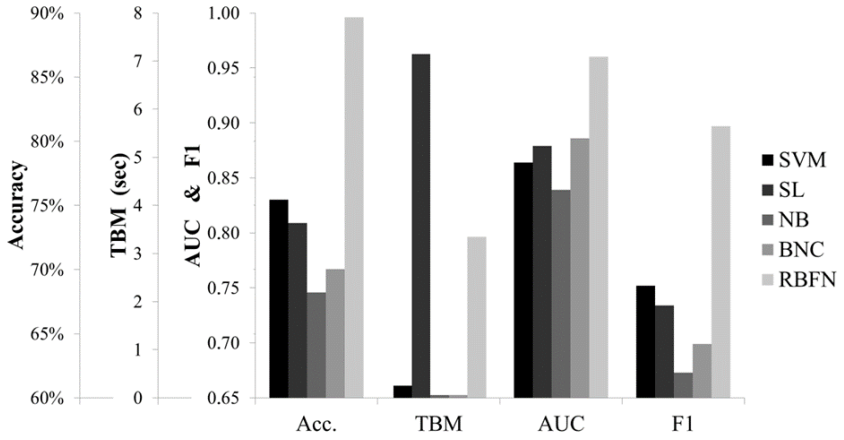

**Fig. 1.** Comparison of the best performances of the five models

As can be seen, the RBFN model outperforms all other models in terms of accuracy, AUC, and F-score. SVM follows as second in accuracy and F-score, but also lags behind SL and BNC in AUC. NB ranks last out of the five models in performance, but is the fastest of all, along with the BNC.

The accuracy of the five models, against the number of keystroke dynamics features used, is shown in Figure 2.

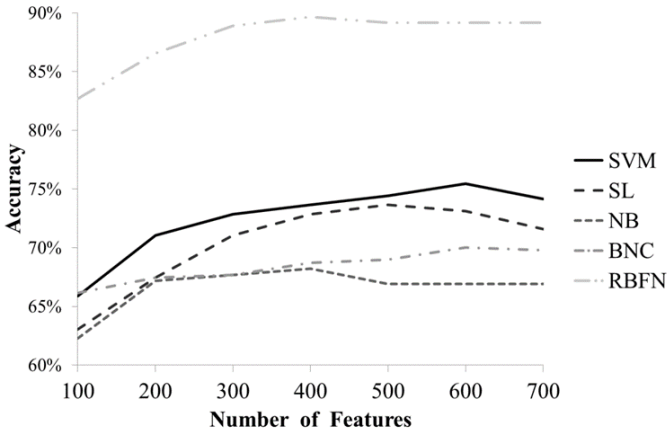


**Fig. 2.** Accuracy of five models on various feature sets

It can be seen from Figure 2 that RBFN presents the highest accuracy against all other models, regardless of the feature set, whereas NB presents the lowest. Also, the performance of BNC has the least dependence on the number of features used, as its accuracy is 68% ± 2% regardless of size of the feature set. One last point is that each of the models achieves the highest precision using less than 700 features. This is a strong indication that a large number of features is not necessarily required to achieve high accuracy, and perhaps of more importance is the use of appropriate features.

## 5 Future Research Directions

Research into the creation of systems that can recognize some characteristics of an unknown user by the way the keyboard is used, has already shown significant results. However, it can be extended in various directions in order to improve the reliability of those systems.

One such direction would be to carry out a new phase of volunteer recording, which will enlarge the existing keystroke dynamics dataset. One of the objectives of this new phase would be to record users of as many different ethnicities as possible, so that the dataset is enriched with logfiles of many different mother tongues, each of which will be adequately represented. The purpose of the study will be to search for

indications as to whether the keystroke dynamics features are independent of the language being typed and, if they are not, the nature of the dependency, and with how much accuracy an unknown user's mother tongue can be predicted. Alternatively, existing datasets such as the one created by Clarkson University [20] can be used.

Apart from the age and the mother tongue of an individual, there are some other characteristics, either acquired or inherent, that can be used to form the user's profile. Among them is gender, for which there are already several studies with very good results. Also, the user's handedness, and their educational level. Creating an integrated system that can extract from an unknown user an accurate set of characteristics, with a very high percentage of success, is the ultimate goal of this research. Indeed, if this is achieved with as little keystroke dynamics data as possible, e.g. a few tens or hundreds of keystrokes, then it could be used as a valuable forensic tool when a digital crime has been committed.

Another interesting direction is to investigate how much a user can modify the way that they type, so as to mislead a recognition system into producing false results. It leads on to the interesting question of whether a user's typing pattern is something superficial, which can be easily disguised, or whether it is more deep rooted, implying that it will inevitably be detected by a suitable system, thus enabling the user profile to be extracted correctly after a certain number of keystrokes.

Finally, another possible avenue of further research is the use of alternative features of keystroke dynamics. As has already been mentioned, in addition to keystroke durations and digram latencies which are widely used, there are many other keystroke dynamics features that could provide much more information to aid in identifying user characteristics. For example, quantitative and qualitative differentiation in pauses during typing may contain some information to enable characteristics classification. Moreover, the preference shown by some users for using a specific key when there are two similar options, such as the "Shift" key, may also contain extra information. The study of the possibility of classifying users with some hitherto underused features, and their combination with those that are widely used, is another possible extension of this work.

## 6 Conclusion

The evolution of technology and the transfer of a large part of human activity to the Internet, including communication, work, entertainment, and education, has also resulted in increased crime in this area. A crime that has very different quality characteristics to that of the physical world, as everyone can hide behind a device and a public network.

This research attempts to identify the age of an unknown user as a part of his/her profile, amongst other characteristics such as gender, handedness, and educational level, according to the way he/she is typing, in order to facilitate a forensic investigation. For this reason, keystroke dynamics features, namely keystroke durations and digram latencies, were extracted from a dataset of 387 logfiles created for this purpose. The process followed was the initial extraction of features, then the selection of

some of them according to the amount of information they include which would aid age classification, and finally the use of some well-known machine learning models for the classification of a user into one of four age groups.

The results showed that the age group a user belongs can be predicted with an accuracy of almost 90%, far exceeding the random prediction accuracy which is 25% (one of four possible choices) and the accuracy level reported in the authors' previous research, which was 66.1%. Therefore, it seems possible to identify some characteristics of an unknown user with high accuracy rates and thereby implement systems that can create profiles of malicious users.

The ability to identify gender, age, and other characteristics of a user may have useful applications beyond forensic investigation. For example, in the field of targeted advertising, where filters are used to determine whether or not a user is exposed to advertising material, there will be much more data at their disposal, resulting in more efficient and less disturbing advertising. Some other applications relate to user convenience, such as automatically filling in certain fields of a form when creating an account, or the suggestions given to a user for things like visiting websites, or joining groups, depending on his/her interests. Finally, an equally important application is that of alerting unsuspecting users to the possibility of becoming a victim of an online fraud. For example, it could be used to warn a young person, who thought that they were communicating with someone of a similar age to themselves, that in fact the person that they were messaging belonged to the age group "46+", with a very high probability.

## Acknowledgement

This work has been partially supported by IDEAL-CITIES; a European Union’s Horizon 2020 research and innovation staff exchange programme (RISE) under the Marie Skłodowska-Curie grant agreement No 778229.

## References

1. Tsimperidis, I., Rostami, S., Katos, V.: Age detection through keystroke dynamics from user authentication failures. International Journal of Digital Crime and Forensics 9(1), 1-16 (2017).
2. Mendoza, D.K.O.: The vulnerability of cyberspace - The cyber crime. Journal of Forensic Science & Criminal Investigation 2(1), 1-8 (2017).
3. Jirousek, R., Shenoy, P.P.: A new definition of entropy of belief functions in the Dempster-Shafer theory. International Journal of Approximate Reasoning 92(1), 49-65 (2018).
4. Zhang, D., Li, S., Wang, H., Zhou G.: User classification with multiple textual perspectives. In: Proceedings of 26th International Conference on Computational Linguistics, pp. 2112-2121. The COLING 2016 Organizing Committee, Osaka, Japan (2016).
5. Culotta, A., Ravi, N.K., Cutler, J.: Predicting Twitter user demographics using distant supervision from website traffic data. Journal of Artificial Intelligence Research 55, 389-408 (2016).

6. Zhang, K., Tan, L., Li, Z., Qiao, Y.: Gender and smile classification using deep convolutional neural networks. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34-38. IEEE, Las Vegas, NV, USA (2016).
7. Chikkala, R., Edara, S., Bhima, P.: Human facial image age group classification based on third order four pixel pattern (TOFP) of wavelet image. The International Arab Journal of Information Technology 16(1), 30-40 (2019).
8. Walther, J.B., Van Der Heide, B., Ramirez, A.J., Burgoon, J., Pena, J.: Interpersonal and hyperpersonal dimensions of computer-mediated communication. In: Sundar, S.S. (ed.) The Handbook of Psychology and Communication Technology, pp. 3-22. John Wiley & Sons, Inc. (2015).
9. Salem, A., Obaidat, M.S.: A novel security scheme for behavioral authentication systems based on keystroke dynamics. Security and Privacy 2(2), 1-11 (2019).
10. Saini, B.S., Kaur, N., Bhatia, K.S.: Position independent mobile user authentication using keystroke dynamics. In: Pandey, B., Khamparia, A. (eds.) Hidden Link Prediction in Stochastic Social Networks, pp. 64-78. IGI Global (2019).
11. Kolakowska, A.: Recognizing emotions on the basis of keystroke dynamics. In: Proceedings of 8th International Conference on Human System Interaction, pp. 75-80. IEEE, Warsaw, Poland (2015).
12. Tsimperidis, I., Yoo, P.D., Taha, K., Mylonas, A., Katos, V.: $R^2$BN: An adaptive model for keystroke-dynamics-based educational level classification. IEEE Transactions on Cybernetics 50(2), 525-535 (2020).
13. Brizan, D.G., Goodkind, A., Koch, P., Balagani, K., Phoha, V.V., Rosenberg, A.: Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. International Journal of Human-Computer Studies 82, 57-68 (2015).
14. Buriro, A., Akhtar, Z., Crispo, B., Del Frari, F.: Age, gender and operating-hand estimation on smart mobile devices. In: Proceedings of 2016 International Conference of the Biometrics Special Interest Group, pp. 273-280. IEEE, Darmstadt, Germany (2016).
15. Roy, S., Roy, R., Sinha, D.D.: ACO-Random forest approach to protect the kids from Internet threats through keystroke. International Journal of Engineering and Technology 9(3S), 279-285 (2017).
16. Pentel, A.: Predicting user age by keystroke dynamics. In: Silhavy, R. (ed.) Artificial Intelligence and Algorithms in Intelligent Systems, pp. 336-343. Springer International Publishing (2018).
17. Guadalupe, T., Mathias, S.R., vanErp, T.G.M. et al.: Human subcortical brain asymmetries in 15,847 people worldwide reveal effects of age and sex. Brain Imaging and Behavior 11(5), 1497-1514 (2017).
18. Osanaiye, O., Cai, H., Choo, K.R., Dehghantanha, A., Xu, Z., Dlodlo, M.: Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. EURASIP Journal on Wireless Communications and Networking 2016(1) (2016).
19. Hu, N.: Using receiver operating characteristic (ROC) analysis to evaluate information-based decision-making. In: Khosrow-Pour, M. (ed.) Advanced Methodologies and Technologies in Business Operations and Management, pp. 764-776. IGI Global (2019).
20. Clarkson University Keystroke Dataset, https://citer.clarkson.edu/research-resources/biometric-dataset-collections-2/clarkson-university-keystroke-dataset/, last accessed 2020/08/30.