S.I. : CSI2017

CrossMark

The probability of predicting personality traits by the way user types on touch screen

Soumen Roy[1] · Utpal Roy[2] · D. D. Sinha[1]

Abstract

Age-group, gender, handedness and number of hands used are the common personality traits of a typist, and identifying such traits can be a key in identifying the person in today's fast world. This particular piece of work is the objective, i.e., an indicative pathway, toward that goal by monitoring and analyzing the way a user types on a touch screen of a smartphone. Study of such traits and analyzing the typing pattern on a conventional computer keyboard has been investigated well. But the conventional keyboard is being replaced with the advent of smartphones with a variety of features, low cost and portability. Therefore, identifying traits through the touch screen is more significant and might be notably beneficial for personal identity prediction and verification. In this paper, we discuss the data acquisition method, classification approach and the evaluation process which are found as more appropriate to discover the trait identities to be used in variety of Web-based applications specifically in the area of e-commerce, online examination, digital forensics, targeted advertisement, age-restricted access control, human–machine interaction, social networks, user identity verification akin to biometrics. Multiple machine learning (ML) methods were used to develop the model, and more suitable and practical evaluation test option—leave-one-user-out cross-validation—was used to check the validity of the proposed model. The efficacy of our approach is illustrated on the dataset collected in the Web-based environment from 92 volunteers. The probability of predicting a user with such traits has also been illustrated here. The study shows timing features of primary keystroke dynamics incorporated with the traits, and the user identification accuracy can be gained up to 17%.

Keywords Biometrics · Keystroke dynamics · Machine learning · Soft biometric · Personality traits · LOUOCV

1 Introduction

Nowadays, common personal trait recognition automatically, basically age-group and gender, is an active area of research. It has tremendous application areas. Identifying these traits

✉ Soumen Roy
  soumen.roy_2007@yahoo.co.in

  Utpal Roy
  roy.utpal@gmail.com

  D. D. Sinha
  devadatta.sinha@gmail.com

[1] Department of Computer Science and Engineering, University of Calcutta, 92 APC Road, Calcutta 700 009, India

[2] Department of Computer and System Sciences, Visva-Bharati, Santiniketan 731235, India

based on morphological biometric characteristics (e.g. iris, face and fingerprint) is common and has been studied well in the literature. It also provides a more reliable and efficient security solution. But, it needs extra effort to align the face, iris or finger correctly. Keystroke dynamics is a behavioral biometric characteristic that can be used for user identification/authentication in a covert environment in recognizing age-group, gender, handedness and hand(s) used. Here, the user may not be aware that their typing pattern is being measured. In this technique, very little or no extra effort is needed to capture that pattern. It can be used to verify the traits continuously beyond the initial decision taken. In addition, no extra hardware requirement is needed, where the built-in sensor is enough to develop the pattern. Nowadays, various sensors added in a smartphone give an opportunity to extract potential features for further development in this domain. Therefore, it is more appropriate to identify the traits based on the way user types on a touch screen of a smartphone. Some of few researchers have put their effort into keystroke

dynamics using a conventional keyboard in desktop or laptop environment to identify such traits. However, in terms of keystroke dynamics in the smartphone, there have been very little published works.

As of now, the performance of the keystroke dynamics system is not 100% accurate since it is a behavioral biometric solution done by observing the changes in typing pattern of the users on a keyboard or touch screen. The main problem with this technique is the way of typing on the touch screen may change slightly over time. This variation may be permanent or temporary depending on various factors like the emotional state of the person, illness, tiredness, etc. It is not the predictable amount of variation so that we can develop a model accurately based on previously called training patterns. Therefore, the global performance of keystroke dynamics is often not good enough. To progress the performance of keystroke dynamics, incorporation of soft biometric traits which can be extracted from primary biometric data as extra features is the new research trend in the domain of keystroke dynamics in order to make the system accurate and time efficient. Due to the fact and importance, a more suitable method is always demanded to predict the personality traits, so that the performance of keystroke dynamics in a smartphone can reach its aim.

Key press and release time of all entered sequences of characters for a single predefined text have been collected as raw data in data acquisition phase. Thereafter, from the raw data, timing features have been calculated. In our study, five different features have been used which are described in Fig. 1. Here, PR-Time indicates the time interval between key press and release time of the same key, RR-Time indicates the time interval between two subsequent releases, RP-Time indicates the time interval between a key release and next key press and PP-Time is the time interval between two subsequent presses and digraph time is the time interval between one key press and second key release. Total five timing features have been used in our experiment for the entered predefined phrase phase "Kolkata." A very simple text has been considered as per the suggestion suggested by the study [1], and the study concluded that choosing simple short text is better to increase the performance of the keystroke dynamics system. In our another experiment [2], we have seen that simple and short daily used words show the impressive results in keystroke dynamics domain.

The way user types on a keyboard contains the timing features which allow recognizing the traits. From the literature, we evidently observed that incorporation of soft biometric traits changes the performance of the biometric system significantly. Soft biometric traits can be extracted from the primary biometric data, generally used to enhance the performance of the biometric system. In this paper, gender, age-group, handedness and number of hand(s) used characteristics have been considered as soft biometrics traits which
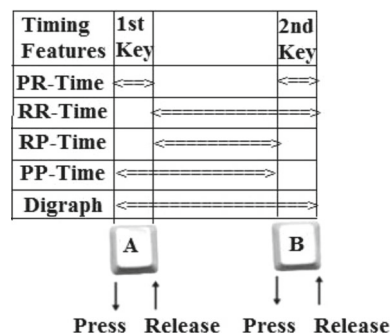

Fig. 1 Timing features in keystroke dynamics used in our models

have been extracted and illustrated. These traits have low user discriminating power but can be used to progress the user recognition performance by fusion of such traits as extra features.

The main research hypothesis of our study is as follows: *Is it possible to recognize the age-group, gender, handedness and number of hands used by analyzing the typing pattern on the touch screen of a smartphone and identifying such traits has any impact on user recognition performance*. In order to experimentally verify the hypothesis, data from the typing pattern on touch screen were collected with labels of age-group, gender, handedness and hands used. To show the impact of incorporation of identifying traits as extra features, the subject information was also collected and labeled.

Most popular and recent ML methods have been selected to train our model as well as to evaluate the respective model. Each model has been trained and evaluated 9200 times for each identifying trait. This study says that if we type "Kolkata" seven times in one session, then our approach can identify the age-group more than 80% correctly where gender, handedness and hand(s) used can be predicted 60, 70 and 80%, respectively. To catch the typing pattern, a predefined reasonable text length is incorporated.

The main objective of this study is to develop a model that can check the personality traits automatically of the users through their typing pattern on the touch screen of a smartphone.

The major goal and contributions of the paper are as follows:

- Develop tools which can be executed as a Web-based application to capture typing pattern on a smartphone.
- Develop a benchmark dataset of 92 subjects collected through smartphone where all the typing records are labeled by their soft biometric traits.
- Develop ML models to identify common personality traits for a single predefined text.
- Introduce a new method of cross-validation which is more appropriate to evaluate the validity of the model.

- Evaluations of model performances have been illustrated in identifying each trait.

The paper is organized as follows. We started by reviewing the overview of previous works on soft biometric on keystroke dynamics in Sect. 2. Then, we present the approach to recognize the gender (male/female), age-group (below 18/18 +), handedness (left/right) and hand(s) (both/single) used in Sect. 3. Section 4 presents the obtained results, it also presents the predicting probability of identifying each personal trait and the last section presents the limitation, scope of works, future works and its conclusion.

2 Related works and analysis

In keystroke dynamics research, researchers spend more time in the data acquisition section rather than addressing challenging issues because this is the most fundamental and essential section of any behavioral biometric system. As a result, various datasets have been produced but none of these is considered to be a balanced one due to many factors ignored during the creation of dataset. Separate datasets have been created with different experimental setups in 37 years of ongoing research, and different evaluation criteria have been controlled on each. Researchers have developed datasets considering only their temporal requirement and method of application; this type of datasets is not so neither potential nor balanced. We need a versatile, well-balanced, reasonable-large dataset to make the study fruitful. It is only because of the lack of standards for data collection and benchmarking. Therefore, it has not been possible to make a sound comparison of the different evaluation processes. In this section, we have explained different approaches, evaluation methods, classification techniques and obtained results in identifying soft biometric traits by analyzing behavior on keyboard or touch screen and also shown the limitation of the proposed works.

The study [3] proposed a novel framework to recognize the gender from keystroke dynamics dataset collected through a computer keyboard. They also have concluded that the inclusion of gender information can enhance the performance of user recognition. All the works they have done are based on the dataset collected through a conventional keyboard. Similar works have been done by the study [4], and they have extracted the gender information along with some other soft biometric characteristics such as age-group ($\leq 30/30+$ or $\leq 32/32+$), hand(s) used and handedness (left/right) of the individual. They have also used a conventional keyboard. Another study [5] also used a conventional keyboard to identify the age-group below 18. But nowadays, keystroke dynamics with the conventional keyboard is going to be obso-

lete where smartphone with various features is available at low cost and became a popular electronic device.

As per the study [6], gender can be reliably predicted by the way user strikes on a touchscreen phone with the accuracy of 64.76%, where the accuracy of 57.16% was achieved on the swipe dataset collected from the same device. Keystroke dynamics characteristics were collected from 42 users where swipe data were collected from 98. Random forest classification algorithm was used to evaluate the model in different cross-validation methods. Leave-one-user-out cross-validation (LOUOCV) and tenfold cross-validation test options were used during the evaluation of the system. Based on the limited dataset, they have concluded that these biometric data can be used to predict the gender reliably.

Another study [7] presented an approach to identify gender using behavioral biometrics pattern generated through smartphones. They have used accelerometer and gyroscope sensor to capture the acceleration and gyroscope data while a user walks in three different modes: normal, fast and slow speed. They have used the data collected from 42 subjects. They obtained the accuracy of 72% to 75% maximum. It is possible to recognize the gender of the user based on the activities on the Web browser as per the study [8]. They obtained 80% of accuracy in tenfold cross-validation test option. In cross-validation, the instances of one subject may be distributed among training and testing sets. Overlapping the instances in both training and testing sets increases the evaluation performance, which is not significant in practice. Uzun et al. [5] showed that it is possible to distinguish the child group and adults through typing pattern on a conventional keyboard, and they obtained the accuracy more than 90% for the simple familiar Turkish text. Multiple classification algorithms were used where SVM (linear) achieved minimum equal error rate (EER) for familiar text. But the performance is not consistent while considering the other text as the study found.

Ancillary information improves the user recognition performance significantly. Some ancillary information like gender can be extracted from the typing pattern as per Giot et al. [3], they obtained the accuracy of 91% to discriminate the gender on GREYC keystroke dynamics dataset created by [9] using libSVM [10] and they also reported that 20% of gain accuracy can be achieved using only gender information as additional feature. Idrus et al. [4] showed that it is possible to identify the gender, age-group ($<30\leq$), handedness and hand(s) used while typing on a conventional keyboard, and they reported the accuracy rate very close to 90%. The objective of extracting soft biometric traits improves the performance of the biometric system. Jain et al. in 2004 [11] obtained an improvement of 5% for fingerprint recognition using ethnicity (Asian/non-Asian), gender (male/female) and height in addition. Ailisto et al. in 2006 [12] decreased the error rate from 3.9 to 1.5% for fingerprint recognition system

using body weight measurement and boy fat percentage for additional information. Jain and Park in 2010 [13] enabled fast search technique in the facial image using freckles, moles and scars as additional information. Li et al. in 2009 [14] achieved the performance improvement of 40% to 50% for face recognition using gender in addition.

As per the literature concerned, a very little work has been studied on the data collected through the smartphone in identifying traits. It also has been observed that some of the few researchers used LOUOCV method in model evaluation. But most of the researchers follow the cross-validation test option in system evaluation. This is not the effective evaluation process as per the study [6].

3 Proposed method

Our first objective was to develop a soft biometric keystroke dynamics benchmark dataset labeled by their gender, age-group, handedness and hand(s) used information. Therefore, a Web-based application was developed to collect the typing pattern of various categories of users in an uncontrolled environment. Then, typing patterns were collected by running the application in a smartphone. Our second objective was to identify the predicting probability of each trait, where some of the ML methods have been used. Our third objective was to evaluate our model with a more appropriate system evaluation test option and check the validity of our approach. The last one is to show the impact and importance of identifying traits in keystroke dynamics user recognition system. The following series of steps have been followed in our study. Experimental setup and process during experiments are described below.

3.1 Data acquisition

For the present purpose, a keystroke dynamics dataset was created through a Web-based application running in a smartphone. During the data acquisition, gender, age-group, handedness and hand(s) used information was collected for each subject. The dataset consists of typing pattern on touchscreen phone from the users belonging to various categories of gender, age-group, handedness and educational qualification which is best suited for the present purpose of identifying traits. Data acquisition method and protocols have been clearly explained in the previous study [15]. The procedures adopted in this study involving volunteers are made following ethical standard.

A specific dataset has been created and selected because it contains the instances by the users belonging to various age-groups (7–18 and 19–65) and gender through the touchscreen environment. In our experiment, samples of typing pattern were collected from 15 children (age below 18) and 77 adults

of 92 subjects in one session with seven repetitions for one text pattern ("Kolkata") through touchscreen mobile device (Moto G Plus 4th Gen., 5.5 inches).

The details of the dataset and sample distribution are described in Fig. 2. The distribution of subjects clearly tells us the class distribution and the diversity of the subjects. Here, all the subjects are daily used smartphone users. It is also clear from the above figure that our dataset is not fully balanced in all categories. But even distribution of classes is sensitive to some of ML techniques. To get the better performance, a balanced dataset is highly needed. In our study, this problem has been taken care.

3.2 Feature extraction and selection

Feature extraction process is carried out for selecting the universal features which are easily available to all the users typing pattern. Basically, in keystroke dynamics, motor behavior, motion behavior and pressure behavior feature subsets are being captured. But motor behavior features are common to all types of keyboard and can be applied to a touchscreen device as well. Although, motion behavior only can be used in a touch environment, where, pressure can be measured with pressure sensitive keyboard. In our experiment, the features were extracted as per the suggestion made by the study [16].

Feature selection is also an important issue in ML research. It enables the method to train faster and reduces the complexity of the model and makes it easy to interpret. It improves the evaluation performance of the model if the proper feature subset is chosen, and it also reduces the overfitting. This part is necessary when the number of features is very large. In our experiment, only the timing features were used. However, accelerometer and gyroscope data are the prominent feature subsets which could be applied with the timing feature subset.

3.3 Preprocessing

Preprocessing step is divided into two subparts. The first is to detect the outliers and clean the data, and the second is normalize the data. In order to improve the classification performance, preprocessing step is necessary since user typing style (raw data) contains noise and outliers. Typing pattern is not consistent due to emotional state, energy level, tiredness, etc. Therefore, the typing style needs preprocessing to detect and remove outliers. Without data preprocessing or data cleaning, classification performance may decline. We have replaced the outliers (value not in between first quartile and third quartile) by the mean value of a specific feature. Then, we normalized the data within the range $[-1, 1]$ with

Fig. 2 Sample distribution of different classes in our dataset
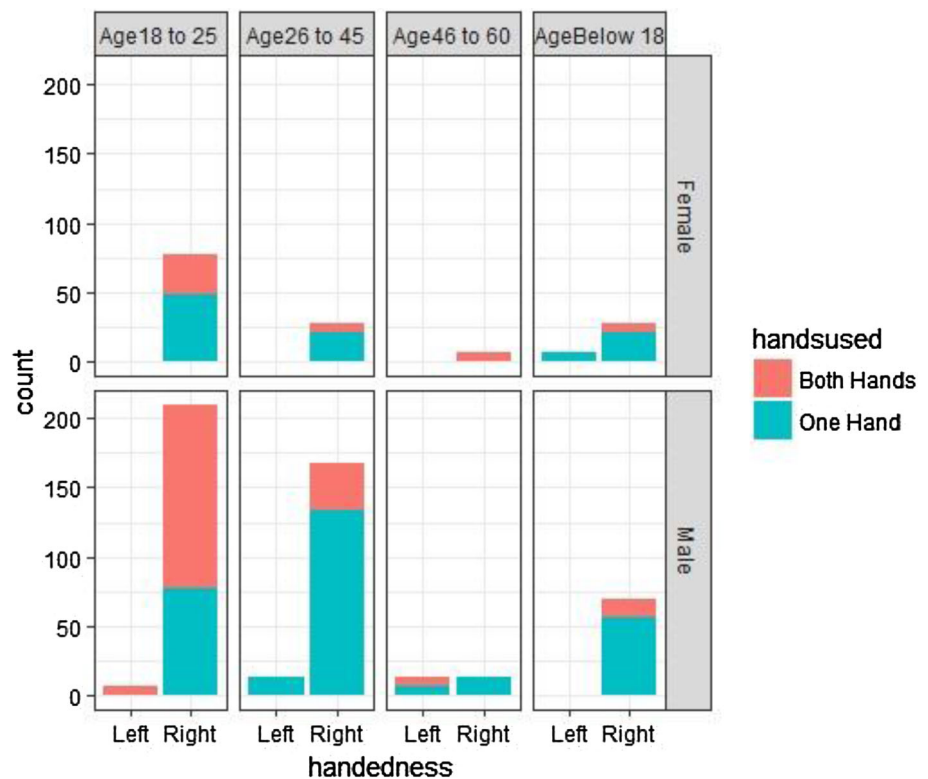
Eq. 1 for faster computing process. Here, x_i is the value which is to be normalized and μ is the mean value of x.

$$x_i = \frac{x_i - \mu}{\max(|x_i - \mu|)} \qquad (1)$$

3.4 Classification and evaluation method

Once the preprocessing step has been completed, classification of age, gender, handedness and hand(s) used is performed based on the similarity and dissimilarities among the different instances. Nowadays, ML methods are common in the biometric pattern recognition system due to the fact that the performances of ML methods are impressive and effective in practice. The variety of ML methods has been applied to our created dataset. Some of the few recent ML methods named Naïve Bayes (NB), rpart, random forest (RF), K-nearest neighbor (K-nn) and C5.0 were used to develop the model and to identify the traits. The performances were recorded with the default parameters of the ML techniques in R [17], defined in the following packages "e1071," "rpart," "caret," "randomForest," and "C50." Before developing the model for identifying each trait, the dataset was divided into training and testing sets subjectwise. For each time, instances of one subject were used as testing set and remaining instances for the other subjects were treated as training set. To make the even distribution of classes in the training sample, the instances of majority classes were selected randomly in each repetition keeping unchanged the instances of minority class in order to balance the training set keeping in mind that some of the ML methods are susceptive to equal arrangement of classes. The undersampling method was used preserving the number of minority class the same and selecting samples randomly from the majority class. This process is executed 100 times for each subject. The model was trained 9200 times for identifying each trait. The classifiers performances were recorded in different performance metrics. The average results were described in this paper with 95% confidence interval (CI).
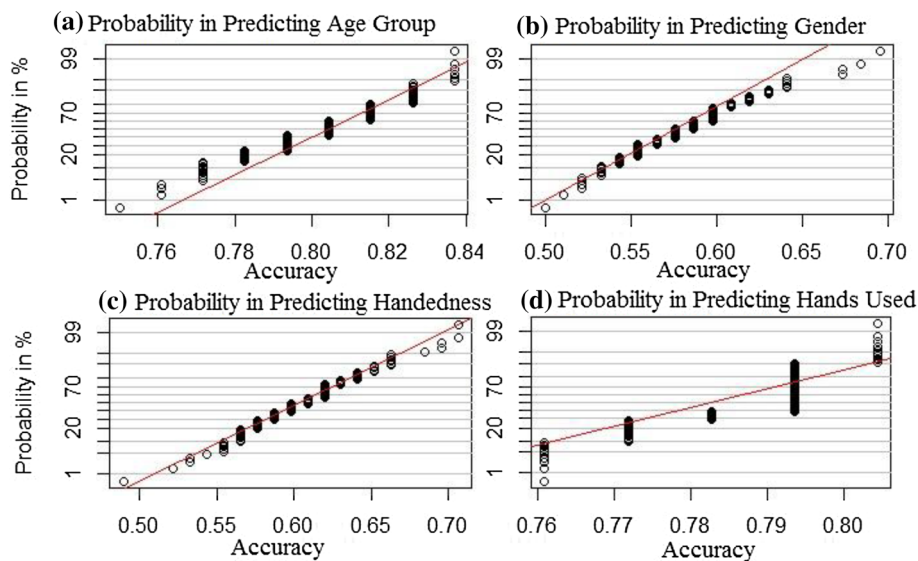
Multiple ML methods were used to train the model, where score fusion of methods was considered to get the better performance in accuracy significantly. In identifying age-group, score fusion of NB and rpart was used, in identifying handedness, score fusion of K-nn and rpart was used as well in identifying handedness, in identifying hands used, score fusion of RF and C5.0 was used intending to get better performance initially.

3.5 Performance metrics

Statistical evaluation of the model can be measured by different metrics. Here, we have used the following four metrics to evaluate the model performance: accuracy, sensitivity, specificity and area under curve (AUC). The following four parameters were used to get such metrics: number of positive classes truly classified (TP), number of negative classes truly

Table 1 CI in identifying soft biometric traits

Traits	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Machine learning methods
Age-group	80.21 ± 0.40	85.13 ± 0.39	54.93 ± 1.16	70.03 ± 0.64	NB and rpart
Gender	58.26 ± 0.76	58.00 ± 1.80	58.34 ± 0.81	58.17 ± 1.00	K-nn and rpart
Handedness	60.59 ± 0.80	37.83 ± 2.39	62.17 ± 0.84	50.00 ± 1.27	NB and rpart
Hands used	78.62 ± 0.59	69.12 ± 0.97	84.46 ± 0.68	76.79 ± 0.63	RF and C5.0

Fig. 3 Probability plot in predicting personality traits

(a) Probability in Predicting Age Group

(b) Probability in Predicting Gender

(c) Probability in Predicting Handedness

(d) Probability in Predicting Hands Used

classified (TN), number of positive classes falsely classified (FP) and number of negative classes falsely classified (FN). Metrics are measured by Eqs. 2–5.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \qquad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \qquad (4)$$

$$\text{AUC} = 1 + \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{FP}}{\text{FP} + \text{TN}} \qquad (5)$$

CI was used to describe the amount of uncertainty associated with the results. CI at 95% is measured by Eq. 6.

$$\text{CI} = \text{Mean value} \pm 1.96 \frac{\sigma}{\sqrt{N}} \qquad (6)$$

Here, σ represents the variation and $N = 100$ represents the number of tries.

4 Experimental results

The obtained results of our proposed approaches are presented in Table 1. The traits and the corresponding ML techniques are depicted in the table with the performance.

It is clear from the table that the accuracies of identifying each trait are impressive for considering a single short text as input. It is also observed that identifying gender is more challenging than identifying age-group.

Figure 3 indicates the probability plot. Here, two components are accuracy and the probability in percentage. In the first figure (a), we can see only 1% chance to get the minimum accuracy below 76% in predicting the age-group but 50% to 80% chances to get 80% to 82% of accuracy in predicting age-group. In the second figure (b), we can see only 1% chance to get the minimum accuracy of 50% in predicting the gender but 50% to 80% chances to get 55% to 65% of accuracy in predicting gender. In the third figure (c), we can see only 1% chance to get the minimum accuracy below 50% in predicting the handedness but 50% to 80% chances to get 55% to 65% of accuracy in predicting handedness. In the fourth figure (d), we can see only 1% chance to get the minimum accuracy of 76% in predicting the hands used but 50% to 80% chances to get 77% to 80% of accuracy in predicting a number of hands used while typing.

In our experiment, we have seen that incorporation of personality traits as soft biometric features increases the performance of the user recognition system in a smartphone. Figure 4 indicates the gain accuracy by fusion of soft biometric information with primary keystroke dynamics characteristics in user recognition system. As per our
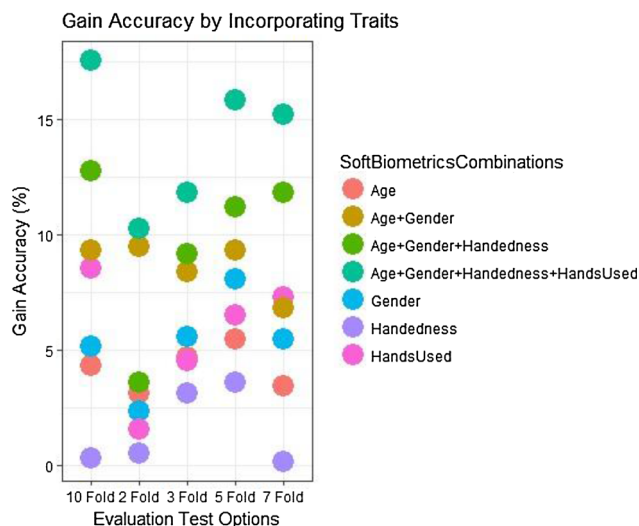
Gain Accuracy by Incorporating Traits

Fig. 4 Impact of incorporation of identifying traits in keystroke

experiment, age-group, gender, handedness and hand(s) used information can improve up to 17% of gain accuracy in a smartphone. If we incorporate all four traits as extra features instead of considering only age or gender information, then maximum gain can be achieved.

5 Conclusions and future work

Age-group, gender, handedness and the number of hand(s) used are the personal identity about the online users which can be predicted through the typing pattern on the touch screen of a smartphone as per the observed results. This paper illustrated the performance of ML methods employed to develop a model in predicting such traits in order to categorize the online users. Identifying traits automatically has many applications that made it necessary to develop a suitable model based on typing pattern for a simple short test which is convenient to use, effective and efficient.

Here, only the timing features have been monitored and considered with a stationary subject. However, various sensors such as gyroscope and accelerometer are available in recent mobile devices to give added opportunities for further exploration with simulated motion. The recent and popular ML method and more significant ML model evaluation test option were used in our experiment, and some challenging issues have been resolved. The experimental results prove the hypothesis. The statistical significance of the present results may motivate some further investigation in this domain with such additional features-level fusion.

The paper first gives an overview of the recent approaches in the domain of revealing traits based on keystroke dynamics. Then, the paper explained the details about the data acquisition protocols and the dataset used in our experiment. Next, it presents the detailed implementations, experimental setup and the approaches. Finally, it shows the obtained results and the statistical significance of those results which may provide some inputs for further research in this domain so that this technique can be effectively implemented in practice in a different mode of applications. Analysis of data, results and inferences here give a primary account toward achievability to the goal with ample scope for further work. There are many parameters prominent and hidden that pose a challenge for future redressing.

References

1. Bartlow N (2005) Username and password verification through keystroke dynamics. Morgantown, West Virginia
2. Roy S, Roy U, Sinha DD (2016) Security enhancement of knowledge-based user authentication through keystroke dynamics. In: MATEC Web of Conferences, 2016, vol 57
3. Giot R, Rosenberger C (2012) A new soft biometric approach for keystroke dynamics based on gender recognition. Int J Inf Technol Manag Spec Issue Adv Trends Biom 11(August):1–16
4. Syed Idrus SZ, Cherrier E, Rosenberger C, Bours P (2014) Soft biometrics for keystroke dynamics: profiling individuals while typing passwords. Comput. Secur 45:147–155
5. Uzun Y, Bicakci K, Uzunay Y (2014) Could we distinguish child users from adults using keystroke dynamics?. https://arxiv.org/pdf/1511.05672.pdf. Accessed 13 Oct 2018
6. Antal M, Nemes G (2016) Gender recognition from mobile biometric data. In: 11th IEEE international symposium on applied computational intelligence and informatics, 2016, pp 243–248
7. Jain A, Kanhangad V (2016) Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings. In: 2016 International conference on computational techniques in information and communication technologies. ICCTICT 2016—proceedings, pp 597–602
8. Kolakowska A, Landowska A, Jarmolkowicz P, Jarmolkowicz M, Sobota K (2016) Automatic recognition of males and females among web browser users based on behavioural patterns of peripherals usage. Internet Res. 26(5):1093–1111
9. Giot R et al (2009) GREYC keystroke : a benchmark for keystroke dynamics biometric systems. In: IEEE international conference on biometrics: theory, applications and systems (BTAS 2009)
10. Chang C, Lin C (2001) LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2:1–39
11. Jain AK, Dass SC, Nandakumar K (2004) Can soft biometric traits assist user recognition? Spie 5404:561–572
12. Ailisto H, Vildjiounaite E, Lindholm M, Mäkelä SM, Peltola J (2006) Soft biometrics-combining body weight and fat measurements with fingerprint biometrics. Pattern Recognit. Lett. 27(5):325–334
13. Park U, Jain A (2010) Face matching and retrival using soft biometrics information forensics and security. IEEE Trans 5(3):406–415

14. Li Z, Zhou X, Huang TS (2009) Spatial Gaussian mixture model for gender recognition. In IEEE 16th international conference on image processing (ICIP 2009)
15. Roy S, Roy U, Sinha D (2018) Identifying soft biometric traits through typing pattern on touchscreen phone. In: Mandal J., Sinha D. (eds), Social transformation—digital way. CSI 2018. Communicaeetions in computer and information science, vol 836. Springer, Singapore
16. Roy S, Roy U, Sinha DD (2016) Comparative study of various features-mining-based classifiers in different keystroke dynamics datasets. In: Smart innovation, systems and technologies
17. RC Team (2017) R: a language and environment for statistical computing. In: R foundation for statistical computing, 2017. www.R-project.org. Accessed 13 Oct 2018